PART I: SUPER-RESOLUTION MICROSCOPY METHOD DEVELOPMENT
PART II: INVESTIGATIONS OF TRANSCRIPTION REGULATION BY CHROMOSOMAL ORGANIZATION IN BACTERIA

by

Christopher Herrick Bohrer

A dissertation submitted to The Johns Hopkins University
in conformity with the requirements for the degree of
Doctor of Philosophy

Baltimore, Maryland
January, 2020

Abstract

Part I: SMLM provides not only high-resolution images of molecular assemblies beyond the diffraction limit but also enables quantitative analysis of the dynamics and compositions. However, challenges in imaging and analysis due to cell geometry, resolution limit, and fluorophore properties impede the full potential of SMLM. To address these challenges, I first developed a single-molecule tracking methodology that minimizes the confinement of diffusing molecules to obtain accurate diffusion coefficients and transition rates. Next, I developed a methodology to improve three-dimensional (3D)-SMLM imaging by directly taking into account the variability of 3D point-spread-functions, which produces superior resolution compared to existing methodologies. Finally, I developed a method to correct for blinking-artifacts. Blinking-artifacts are caused by repeated localizations of the same fluorophores, which distort images and produce false nanoclusters. I derived a method to find the "ground-truth" of the underlying pairwise distribution without any additional calibration. This ground truth enables me to identify the true underlying spatial distribution of molecules in the SMLM image, solving a problem that has long persisted in the field.

Part II: It is well established that chromosomal organization dramatically influences transcription, but the underlying mechanisms remain elusive. We hypothesize that supercoiling constrained by the chromosomal topology has an effect on transcription rate and hence coordinates expression within the same topological domain. To examine this hypothesis, I developed a theoretical model to account directly for the buildup of supercoiling due to transcription in a DNA-loop. To investigate how the topology of the chromosome influences transcription further, I then developed the first *in vivo* assays to manipulate the formation of a "large" chromosomal DNA topological domain in *E. coli* cells to examine transcription activity of multiple genes enclosed in the domain. My experiments showed that domain formation decreases expression levels of genes both inside and outside the domain — demonstrating a "long-range" cis-regulatory mechanism due to the "architecture" of the chromosome within bacteria. Finally, using quantitative SMLM, we investigated how "large-scale" chromosome organization affects the spatial organization of RNA-polymerase (RNAP). We discovered RNAP clusters engaged in active ribosomal RNA synthesis; whose organization is "driven" by the chromosomal organization.

Primary Readers and Advisors:

Jie Xiao
Professor
Department of Biophysics and Biophysical Chemistry
Johns Hopkins School of Medicine

Elijah Roberts
Assistant Professor
Department of Biophysics
Johns Hopkins University

Declaration

To my wife, Shannon

Acknowledgements

I would like to start by thanking my advisors/mentors, Dr. Jie Xiao and Dr. Elijah Roberts, who have both tremendously impacted the way I think about and conduct science. I was extremely blessed to have received training in both theory and experiment and to be within an environment where I was able to investigate what I found interesting. Both of my advisors were always extremely supportive in all aspects of my life and I would not have succeeded without them. I would also like to thank Jie for her help with Theodore, where she really helped my family through a most difficult time. I would also like to thank my thesis committee, Dr. Barrick, Dr. Johnson, Dr. Wu, Dr. Myong and Dr. Ha (Alternate) for the support over the years.

Another person who I would consider a mentor is Dr. Xinxing Yang, who taught me so many numerous things it would be impossible to list them all. His love for nature is truly contagious and he is one of the greatest pure scientists this world knows (Even when we have different opinions about statistics — I'm sure one of us will come around eventually).

Next, I would like to thank all my fellow lab mates. First I would like to thank Dr. Jason Lyu and Lior Shachaf for constantly talking about how their children are doing and allowing Little Christopher to play with them (It is always nice to be able to relate to others). Second, I would like to thank my bay mate Joshua McCausland for entertaining my humor and cutting down on the puns when I was present — it was very nice to have someone read and discuss books of interest. Third, I would like to thank Dr. Gina Wang for bringing humor to the lab and always being a "straight shooter." Finally, I would particularly like to thank my friends Dr. Xiaoli Weng and Dr. Kelsey Bettridge for working and daydreaming with me, it was truly a pleasure to work so closely with you two.

During my time at Hopkins I also got to train/mentor quite a few individuals, which was a great deal of fun and I look forward to their futures, as all of them are extremely talented. I would like to specifically thank Max White who helped me get an assay going to investigate how DNA loop formation influences transcription, and Yuncong Geng and Nicolas Yehya who will continue this work (I absolutely loved being around you two). I would also like to specifically thank Yiwen Zhang for showing me that two very different people, who think very differently, can become the best of friends.

I also received a great deal of support outside of the lab during my time at Hopkins. I have always been troubled with being away from my family and my children not always getting to see their relatives whenever they please. Fortu-

nately, the Nevin family (including Laura, Brian and Kelly) was always there; buying the kids toys, inviting us to their family holiday get togethers and providing the finest food (Thank you, you are all truly wonderful people). My best friends Dan Frost, Jon Howell-Day, Henry Lessen and Ryan McQuillen, were instrumental in so many ways in maintaining not only my mental state, but the well being of my family, I cannot thank these four enough (Especially Ryan and Henry, these two were with me through it all: watching my kids when I was sick, staying up late with my children when Shannon and myself could not be there, being Kit's best friends, thank you.) Also, I would like to thank Mere Peck and Iliad for the many early morning runs and coffee.

I also want to thank the people who are truly responsible for making me who I am. I would like to start with my Grandmother (Betty Wesner) and my Uncle/Godfather (Mark Wesner) for always being there for me. To my siblings Nicki (thank you for calling me almost every day, as it was nice to feel connected again), Brad and Austin, you are truly my greatest and closest friends — thank you for everything you do for me and my kids (Also, thank you Julie and welcome to the family, good luck!). I would also like to thank my role model, my father, who is always in a good mood, up for a run, has the most amazing work ethic and won't hesitate to drive 12 hours just to say "hi", thank you. I owe the greatest amount of thanks to my mother (especially in regard to my education), I pray to have your drive, patience, determination and love when raising and educating my own children, thank you for never giving up on me.

Finally, I would like to thank Theo, Kit and my wife Shannon, you are everything to me and I am truly blessed to have you in my life.

Contents

List of Tables

List of Figures

Chapter 1

Complex Diffusive Behavior Within Bacteria

1.1 Introduction

Diffusion is the consequence of a particle randomly colliding with other particles in its surroundings. The diffusion speed, directionality, and trajectory of a particle contain rich information about how the particle interacts with its surroundings, offering an invaluable window to examine molecular interactions in live cells.

In bacterial cells, random diffusion is sufficient to allow molecules to reach their desired target sites efficiently because of the small cellular volumes. For example, a protein molecule with a diffusion coefficient D of 1 $\mu m^2/s$ can sample the entire cytoplasm in $\sim$ 100 ms. In contrast, eukaryotic cells have volumes that are three orders of magnitude larger and simple diffusion is no longer sufficient. Directional motor proteins such as kinesin and myosin are hence required to deliver molecules to different cellular addresses. Since diffusion is a major mechanism behind how molecules find their "place" in bacterial

cells, it is vital to understand the characteristics of diffusion in the different compartments of bacterial cells. Here we present a critical summary and evaluation of commonly used methods and analyses to probe complex diffusive behaviors observed in bacterial cells, with a major focus on single-molecule tracking (SMT). We then elucidate various diffusion dynamics with specific examples in the bacterial cytoplasm, nucleoid, and the membranes.

1.2 Common methods to characterize diffusion in bacterial cells

Commonly used methods to characterize diffusion in live cells are fluorescence recovery after photobleaching (FRAP), fluorescence correlation spectroscopy (FCS) and single molecule tracking (SMT). Here we briefly describe FRAP and FCS, and then discuss SMT in depth, due to its wide use and vast potential in probing diffusion in bacterial cells.

Fluorescence recovery after photobleaching (FRAP)

In FRAP, a focused laser is used to photobleach a small region of a cell containing fluorescently labeled molecules, and subsequently the fluorescence recovery of the region is monitored. Depending on the diffusion speed, diffusive mode, and the geometry of the selected region and cell, the FRAP curve can be fit to specific models to extract diffusion coefficients and kinetic rates associated with particular molecular interactions [1]. As an ensemble method, FRAP is relatively simple to implement; the apparent FRAP rate serves as a straightforward measure to allow comparison of the same system under dif-

ferent conditions even in the absence of a specific model. Therefore, FRAP has been widely used in diffusion studies. However, one should be aware of the limitations of using FRAP to extract quantitative parameters such as diffusion coefficients and kinetics. These values are ensemble-averaged means pertinent to and only valid in specific models. Finally, FRAP is unable to depict heterogeneous diffusion properties of molecules limiting its use in terms of determining diffusive behavior [2].

Fluorescence correlation spectroscopy (FCS)

FCS is an ensemble methodology that monitors the fluctuations of fluorescence within a small region to determine many different parameters, including the diffusion coefficients. The mechanism of the technique is that the fluctuations in fluorescence are due to molecules moving into and out of the illuminated region, allowing the dynamics of the system to be quantified. FCS often includes calculating the autocorrelation function and then fitting it to specific models to extract the desired parameters. When compared to FRAP, the theoretical interpretation of the data is really quite similar and the same limitations exist in regards to quantifying the diffusive behavior [3].

Single-molecule tracking (SMT)

SMT is an *in. vivo* method where one follows the movement of individual molecules (or particles in some cases) labeled with fluorophores to determine how the molecule interacts with its surroundings and potential targets. Because of the real-time and single-molecule nature of the method, SMT allows one to identify not only the molecule's diffusive mode and diffusion coefficient,

but also the population heterogeneity and *in vivo* kinetics of switching between diffusive states, which are often indicative of specific molecular interactions.

1.3 Practical concerns of SMT

A successful SMT experiment requires a few critical parameters be within an optimal range. These parameters are the single molecule signal to noise ratio (SNR), the length (L) and number (N) of SMT trajectories. These parameters have large influences on the theoretical limitations of quantifying diffusion coefficients and in being able to quantify the diffusive behavior with different forms of analysis, discussed throughout [4].

SNR

The first parameter, SNR, is often defined as the ratio between the number of photons emitted by the fluorophore and the cell's autofluorescence background. The SNR dictates how well one can determine the position of a molecule at each time point, i.e., the precision at which the molecule can be localized. This concept is the same as the localization precision from single-molecule localization superresolution microscopy (SMLM). In live bacterial cells, the SNR for commonly used fluorescent proteins and organic dyes are sufficient for a localization precision of $\sim 10 - 30$ nm [5, 6, 7]. When the SNR is low due to a short camera exposure time or a high cellular autofluorescence background, individual displacements along the SMT trajectory cannot be determined accurately, leading to large uncertainties in determining the corresponding D [8, 9, 10, 11, 12]. As such, metrics used to quantify the diffusive dynamics of molecules [13, 14, 15, 16, 17] are often distorted, making the interpretation

of the data difficult and complex [13]. For instance, when the SNR is low, the mean squared displacement (MSD, discussed in detail below) can show sub-diffusive behaviors at short timescales, even when the diffusion is purely Brownian [18]. Additionally, the quantified diffusive states and corresponding kinetic switching rates can be ill-defined, due to the low confidence in defining D values along a trajectory [9, 10, 4].

Trajectory length (L)

The second parameter, L, describes how long in time a molecule can be tracked. In practice, L is limited by the time a fluorophore remains fluorescent before it photobleaches. Due to the stochastic nature of photobleaching (photobleaching time is usually exponentially distributed [19, 20]), only a small portion of all SMT trajectories are relatively long. As such, a large number of total SMT trajectories are often required to obtain a sufficient number of long SMT trajectories.

Long SMT trajectories are vital to determine if the diffusive behavior of molecules is ergodic, if there is dynamic heterogeneity along single trajectories, and if the molecule transitions between different diffusive states. Here ergodicity refers to whether the average behavior across all molecules equals the behavior of individual molecules over long periods of time, which can be used as a metric to discriminate between different models of diffusion [21, 22]; dynamic heterogeneity means the diffusion coefficient of an individual molecule varies through time or space [23]; transition kinetics refer to the rates of a molecule switching from one to the other diffusive state characterized by distinct D's.

How long is long enough for SMT? In ensemble kinetic measurements of chemical reactions, a rule of thumb is to monitor the reaction for at least four reaction halftimes in order to determine the rate constant accurately. The equivalent should be applied to SMT as well. For example, for a transition rate of 1 s^{-1}, the minimal average trajectory length should be at least $\sim 4s$ long in order to capture a sufficient number of transition events. In practice, SMT tracking trajectories should be even longer in order to observe the two different states before and after the transition with confidence. In theory, one can also obtain a large number of shorter SMT trajectories (>10,000) and analyze the data using statistic methodologies to extract the kinetic information [9, 8, 10]. These statistical methods often require additional assumptions about the kinetic rates and steady states, and hence need to be carefully evaluated.

How to achieve high SNR and obtain long trajectories

To achieve a high SNR, the key is to use bright fluorophores. To obtain long trajectories, the key is to use photostable fluorophores. Bright, red-colored organic fluorophores such as the newly developed JF646 dye [24] (now commercially available) in conjunction with Halo or SNAP tag [25, 26] satisfies both requirements and is the best choice for SMT. The unique, rigid fluorophore structure of JF646 ensures high fluorescence quantum yield and low photobleaching quantum yield, and the lengthened conjugation plane allows red-shifted excitation at 647 nm, which avoids the autofluorescence background that usually comes from flavin proteins [27]. Halo- or SNAP-ligand modified JF646 is membrane permeable (even for Gram negative bacteria such as *E. coli*) and can be directly added into cell's growth medium and subsequently

washed for live cell labeling.

In the event that the Halo/SNAP-JF646 labeling system or its alike is not feasible (for example, the fusion protein is not functional or there is a high level of nonspecific dye binding), other strategies can be employed. Fluorescent proteins (FPs) usually tolerate fusions well and do not require the addition of exogenous fluorophore, simplifying sample preparation. In our experience, the red-colored fluorescent protein TagRFP-t [28], even though not comparable to JF646, is sufficiently bright and is the most photostable when compared to the other FPs we tested (EGFP, EYFP, mCherry, mNeoGreen, mEos3.2, and PAmCherry) [29, 30, 31, 32, 33, 34]. If other less bright or photostable fluorophores are the only option, one could try to [1] minimize cellular autofluorescence background by avoiding the green-colored fluorophores and by growing cells in defined (such as M9 or EZRDM) instead of complex (such as LB) media; [2] conduct multiple rounds of SMT experiments in which the dark interval between adjacent imaging frames is systematically varied so that trajectories of different dark intervals can be computationally stitched together to cover longer time scales [35].

Trajectory number (N)

A final requirement for a successful SMT experiment is to obtain a sufficient number of trajectories. As with any single-molecule experiment, diffusive trajectories of individual molecules are inherently stochastic and therefore a large sample size is needed to quantify and account for the fluctuations. Generally speaking, to obtain a single mean diffusion coefficient D a hundred trajectories with an average length of at least five to ten tracking frames may be

sufficient. If there are multiple populations with different D, a few hundred to a thousand trajectories are necessary to separate the different populations. To extract kinetic rates, greater than 10,000 trajectories may be needed (see below for more details).

The above requirement demands collecting as many trajectories from as many single cells as possible. However, SMT also requires a low labeling density in single cells so that individual molecules can be spatially isolated. A low labeling density can be achieved by carefully tuning the fusion protein's expression level using repressible promoters and/or low copy plasmids, so that on average in one bacterial cell there is only one or two fluorescent molecules. Such a low expression level leads to a low imaging throughput since the majority of cells would not have any expressed fluorescent molecules. Furthermore, the expression level is often difficult to control experimentally due to the leakiness of most prokaryotic promoters.

One way to bypass this experimental difficulty is to use the Halo/SNAP-JF dye labeling system. The fusion protein can be expressed normally in cells, but the concentration of the dye can be tuned at will so that only a small percentage of fusion protein molecules is labeled to allow SMT. This strategy, however, still does not circumvent the issue of low data throughput, since one can only obtain on average one or two trajectories per cell (for bacteria).

The ideal strategy is to use a photoactivatable fluorophore that is not fluorescent unless activated [36, 37, 38, 34]. A fusion protein could thus be 100% labeled with a photoactivatable fluorophore but remain nonfluorescent; only upon a low dose of activation light one or a few molecules are stochastically

turned on to be tracked. After they are photobleached, new molecules are turned on, allowing continuously SMT of many molecules in the same cells. Commonly used photoactivatable fluorophores include mEso3.2 [33], PAm-Cherry [34] and the new photoactivatable JF dyes [39], but none of their photochemistry properties are as good as the stable JF646, and their applications in SMT are still relatively limited. Photoactivable JF dyes have been developed [39], but their low activation rates require further optimization for SMT. Furthermore, continuous photoactivation using high energy light (405 or 488 nm) can cause photodamage of cells, limiting the number of trajectories one can continuously collect from individual cells.

1.4 Data analysis and interpretation of SMT

SMT trajectories can be analyzed multiple ways depending on what quantitative information one wishes to extract. Commonly used analyses include mean squared displacement (MSD), cumulative displacement distribution function (CDF), velocity autocorrelation function (VAF), and Hidden Markov Model (HMM). Below we describe each analysis and what information can be determined independently and collectively from these analyses.

Mean squared displacement (MSD)

The mean squared displacement (MSD) is the most commonly used metric to estimate the apparent diffusion coefficient D, which helps quantify the diffusion mode of single molecules (see the section: Commonly encountered diffusion mechanisms below). The ensemble-averaged MSD is calculated by taking the squared distance a molecule travels for a certain time and then averaging over

all molecules:

$$MSD(t) = \frac{1}{n}\sum_{i=1}^{n}(x(t) - x(0))^2 \tag{1.1}$$

where $x(t)$ is the coordinate of the molecule at time t and n is the number of trajectories. Note that in all experimentally measured MSD curves, the square root of the y-axis intercept, or the apparent MSD value when $t = 0$, indicates the uncertainty in determining a molecule's position, hence serving as a useful indicator to estimate experimental localization precision. Because different SMT trajectories have different lengths, to ensure that each molecule contributes equally to the final MSD curve for each t, a common practice is to select trajectories that have a minimal length and truncate longer trajectories to the minimal length.

For individual trajectories, the time averaged MSD of each trajectory is computed using the following equation:

$$MSD_\tau(t) = \frac{1}{T-t}\sum_{\tau=0}^{T-t}(x(t+\tau) - x(\tau))^2, \tag{1.2}$$

where T is the total time of the individual trajectory and τ ranges over all possible values up to $T - t$ based on the time interval of the SMT experiment. If a system is ergodic, the $MSD_\tau(t) = MSD(t)$ and it can be used to discriminate between different modes of diffusion, Fig. 1.1B. Note that a combination of the two can be used to examine ergodicity if trajectories are not of sufficient length [22]:

$$MSD_{\tau}^{avg}(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{T-t}\sum_{\tau=0}^{T-t}(x_i(t+\tau) - x_i(\tau))^2. \quad (1.3)$$

CDF

As mentioned above, dynamic heterogeneity means that D values of individual molecules vary through time and/or space. Dynamic heterogeneity can exist simply because the molecule of interest has multiple diffusive states depending on its interactions with other molecules. For instance, in *E. coli*, RNA polymerase (RNAP) molecules exhibit a D of $\sim 1\mu m^2/s$ in the cytoplasm, $\sim 0.4\mu m^2/s$ in the nucleoid, and $\sim 0.1\mu m^2/s$ when bound to chromosomal DNAs ([40] and see section: Diffusion of DNA binding proteins). These different D values indicate different modes of DNA interactions of RNAP, with the slowest one most likely bound to DNA, the fastest one freely diffusing in the cytoplasm, and the intermediate one interacting with the nucleoid nonspecifically. Dynamic heterogeneity can also result from the molecule experiencing different local environments within the cell [41]. When a "continuum" of heterogeneity is observed, the varying diffusive properties could be due to the local environment changing with time or the molecule moving to a different environment.

One useful way of examining whether there are multiple diffusive populations is to inspect the displacement distribution. For a single population with 1d-Brownian motion (random collisions of the molecule with other molecules within the medium), the displacement distribution for a molecule to move a distance x away from the origin in the time interval t follows a normal distri-

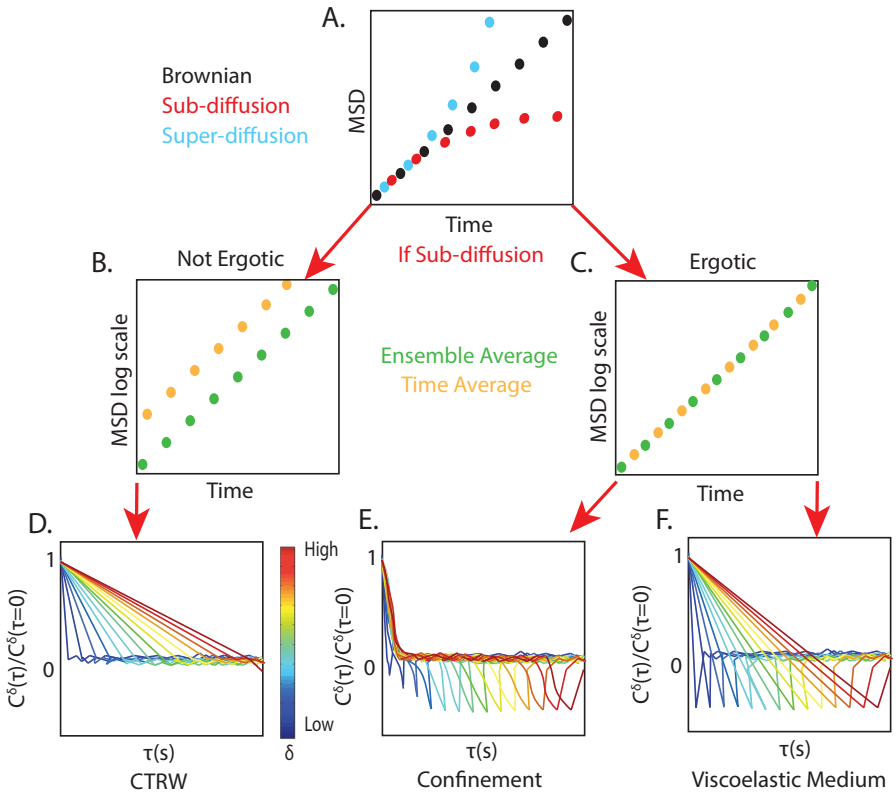

Figure 1.1: Characterizing different types of diffusion. A. The MSD (linear scale) for the three catagories of diffusion. B. The MSD (on a log scale) for a non-ergotic system. C. The MSD (on a log scale) for an ergotic system. D. The VAF that results for a Continuous Time Random Walk model of diffusion. E. The VAF that results due to diffusion with confinement. F. The VAF that results from diffusion within a viscoelastic medium. (D-F: color indicates δ)

bution

$$p(x,t) = \frac{e^{\frac{-x^2}{4Dt}}}{(4\pi Dt)^{\frac{1}{2}}}, \tag{1.4}$$

with a characteristic diffusion coefficient D, that is dependent upon the size of the molecule, the temperature and the viscosity of the medium. The corresponding cumulative distribution function (CDF), can also be fit to a single exponential function to extract D. When the displacement distribution or CDF cannot be described adequately with a single population, the sum of multiple Gaussian or exponential functions with different D values and respective population fractions can be used [42]. The CDF for two populations of diffusing molecules can be fit with the following equation:

$$CDF(r,t) = 1 - \alpha \times e^{\frac{-r^2}{4D_1 t + 4\sigma^2}} - (1-\alpha) \times e^{\frac{-r^2}{4D_2 t + 4\sigma^2}} \tag{1.5}$$

where D_1 and D_2 are the diffusion coefficients of the two diffusion populations, α accounts for the fraction of each population, r is the radial distance and σ is the localization precision of the experiment. Once different populations with characteristic diffusion coefficients are identified, one can analyze each population's behavior as described below.

Velocity autocorrelation function

If one wishes to characterize the diffusive behavior and dissect the mechanisms behind it, the velocity autocorrelation function (VAF) is an important tool [13].

The function identifies the correlation in the velocity of a molecule at different timescales and allows one to distinguish between different diffusive processes and is particularly useful for sub-diffusion [13]. The function is defined by the following:

$$C_v^{\delta}(\tau) = \langle \vec{v}(t+\tau) \cdot \vec{v}(t) \rangle, \tag{1.6}$$

where

$$v(t) = \frac{1}{\delta}([\vec{R}(t+\delta) - \vec{R}(t)]. \tag{1.7}$$

Here $\vec{R}(t)$ is the position vector of the molecule at time t and $\langle \vec{v}(t+\tau) \cdot \vec{v}(t) \rangle$ is the mean dot product averaged over all trajectories. The values of δ and τ are varied across all possible time intervals of the trajectories. As we will describe more in detail below, specific characteristics of the VAF are indicative of different diffusion modes and when combined with other analyses, it is often possible to delineate the underlying diffusion mechanism.

HMM

In most cases where there exist multiple "well defined" diffusive populations, it will be of interest to identify the transition kinetics between the different states of the molecule. The ability to monitor the change of a molecule's diffusive behavior in real time, which reflects its interactions with targets without perturbing the system is one of the most powerful benefits of SMT [9, 8, 10]. For SMT, to extract the kinetic rates between states, the individual states must have different diffusion coefficients and the displacement distributions need to be known [8, 10]. One can then use likelihood and Bayesian approaches to

quantify the transition kinetics of the system by fitting to a hidden markov model (HMM), see Das et al. for further details [8]. As of yet, the field has not yet determined a methodology for accounting for non-Brownian diffusion, though methodologies are beginning to take the processes responsible for non-Brownian motion into account [9, 11]. One such approach will be discussed in the following chapter of this thesis.

1.5 Commonly encountered diffusion mechanisms

Brownian Motion

Brownian motion, in which a molecule randomly collides with the surrounding molecules, is the most common and the simplest diffusion mechanism. The MSD plot of a SMT experiment is a straight line, with the slope of the line providing the diffusion coefficient (Fig. 1.1A). The corresponding single-step displacement distribution or CDF can be well described by a Gaussian or a single exponential function. Additionally, due to the fact that all displacements are independent of each other, the VAF decays to zero for all $\tau \geq \delta$. Note, if the experiment has a low localization precision, the VAF will show a negative peak for $\tau = \delta$, at low values of τ, which approaches zero as τ increases.

Anomalous Diffusion

Any type of diffusion process that does not result in a linear MSD is considered anomalous. There are two types of anomalous diffusion, sub-diffusion and super-diffusion. Most of the time, anomalous diffusion has an MSD that scales with time to an exponent, $MSD = 4Dt^{\alpha}$. (For sub-diffusion $0 < \alpha < 1$ and for super-diffusion $\alpha > 1$) Examples of the MSDs for both types of anomalous diffusion are shown in Fig. 1.1A.

Super-diffusion usually results from directional movement of molecules and is rare in bacterial cells, as they do not have linear motor proteins such as kinesin or myosin. However, directional movement of cytoskeletal proteins such as MreB [43], cell wall remodeling enzymes PBP2 and PBP3 [44, 45, 46] and segregating plasmid DNAs [47, 48] have been observed. The VAF of super-diffusion will show positive values across large τ values as the directionality of individual displacements is highly positively correlated [49].

Sub-diffusion is commonly observed in bacterial cells and can result from a number of different mechanisms. A first step to differentiate different diffusion mechanism is to compare the exponent value of the MSD curve with what would be expected from the different diffusion models, as what was done previously on the diffusion of chromosomal DNA segments and mRNA molecules [14]. However, because different sub-diffusion processes can result in similar MSD curves, other metrics are normally needed to support specific models. Below we focus on a few models pertinent to diffusing molecules in bacterial cells.

Diffusion with Confinement (sub-diffusion)

The most common mechanism behind sub-diffusion in bacterial cells is confinement, which results from diffusion in a finite space. With confinement, the size of space a molecule can explore is limited and the MSD reaches a plateau at long time scales, causing the MSD to scale with an exponent $\alpha < 1$. The value of the plateau can be used to extract the size of the confinement zone, which corresponds to the finite size of space where the molecule could freely diffuse [50]. For 1d diffusion (along x) within a box of length L_x, the $MSD_x(t)$ follows:

$$MSD_x(t) = \frac{L_x^2}{6} - \frac{16L_x^2}{\pi^4} \sum_{n=1(odd)}^{\infty} \frac{1}{n^4} exp[-\frac{1}{2}(\frac{n\pi 2D_x}{L_x})^2 t] \tag{1.8}$$

Here it can be seen that as $t \to \infty$ the MSD will asymptotically approach $\frac{L_x^2}{6}$, which defines the value of the plateau. If diffusion is Brownian, sub-diffusion caused by confinement will still appear Brownian at short time scales before the molecules can experience the barriers. Therefore, the single-step displacement distribution will still be Gaussian. At long time scales, the displacement distribution or CDF will deviate from that expected from Brownian motion.

The characteristics of confinement can be quantified using VAF. Confinement results in an "anti-persistent" behavior, in which a molecule is reflected off of the barrier and returns to its previous position. The resulting VAF $C_v^{\delta}(\tau)$ shows a small negative peak or a zero at small δ and τ (due to the molecule not having time to experience the barriers) and then develops into a large negative peak as δ and τ increase (the barrier reflects the molecule, leading to

the negative velocity relative to the previous velocity), Fig. 1.1E.

Confinement often leads to difficulties in identifying the true diffusive behavior of molecules. For instance, confinement eliminates long timescale correlations in the VAF [13] and reduces D values, and hence leads to mis-identified diffusion modes and states, creating error in the associated kinetic rates [9].

To limit the amount of confinement in rod-shaped bacterial cells, it is a common practice to take the displacements along the long axis of the cell, as it introduces less confinement when compared to the short axis of the cell due the longer length [9, 51, 10, 14]. However, this practice eliminates a significant amount of data, leading to less accurate determination of D and transition kinetics of a system. Bohrer et al. developed an algorithm, termed Single-Particle tracking Improvement with Confinement Error Reduction or SPICER, to selectively incorporate the displacements along the confined dimension of the cell by quantifying the distance of a molecule to the barrier that is needed to minimize the effects of confinement. The new algorithm significantly improves the accuracy in determining both the D values of different diffusive species and also the associated kinetic transition rates of the systems [9].

Diffusion near a liquids glass transition (sub-diffusion)

Another mechanism of sub-diffusion can be due to a disordered/heterogeneous medium [52]. For instance, it is well known that diffusion deviates from Brownian motion in amorphous solids [53, 54]. Interestingly, the bacterial cytoplasm has been reported to have "glass like properties" and changes from liquid-like to solid-like in a metabolism-dependent fashion [22].

The MSD curve of molecules diffusing in a glass-forming liquid has three

distinct characteristics: [1] At short timescales the MSD displays a linear relationship, indicative of free diffusion; [2] at intermediate timescales, the MSD approaches a plateau due to the molecules being trapped in "cages" formed by the relatively immobile solvent molecules; and [3] at long timescales the cages rearrange allowing the molecules to escape, leading to normal Brownian motion within the MSD [54].

Additionally, diffusion within a "glass-like" medium is non-ergodic, meaning that the average MSD over all trajectories does not equal the average of individual trajectories (over time) (Fig. 1.1B) [55]. The non-ergodicity is the result of the medium having an infinite phase space, in that there is an infinite number of ways to create local cages and unique arrangements of molecules. With the infinite phase space in mind, a medium that is approaching its glass-like transition will have certain areas displaying glass-like properties while others display fluid-like properties, leading to heterogeneities in the diffusion modes of different molecules within the same cell. This mechanism results in individual cells having more than a single population of diffusing molecules, some confined to cages and others freely diffusing [22].

Finally, molecules diffusing in a medium approaching its glass transition also exhibit the anti-persistent behavior [22, 54]. The anti-persistent behavior is exemplified by a strong negative correlation between adjacent displacements. For adjacent displacements there is a strong linear dependence for the magnitudes of adjacent displacements (in the direction of the first displacement) up to the "cage size" of the medium (Fig. 1.3B) [54]. The anti-persistent behavior arises because molecules are reflected by the cage barriers (similar to that

of confinement), causing the molecules to return to their previous positions. However, analyzing the anti-persistent behavior using the relatively simple negative correlation of adjacent displacements instead of the VAF cannot examine the effects of viscoelasticity, localization precision and confinement, as we further discuss below [13, 54, 22].

Diffusion within a viscoelastic medium (sub-diffusion)

In Brownian diffusion, each step a molecule takes is independent of the previous steps. In other diffusion modes there may exist temporal correlations throughout an individual trajectory, which are thought to be a hallmark of complex systems containing many interacting components. Temporal correlations themselves lead to anomalous diffusive behavior [56]: positively correlated subsequent displacements lead to super-diffusion, whereas all other types of temporal correlations produce sub-diffusion.

One common mechanism that leads to temporally correlated sub-diffusion is diffusion within a viscoelastic medium. For example, the diffusive motion of bacterial chromosomal loci in the cytoplasm has been modeled as a polymer within a viscoelastic medium; the viscoelasticity of the fluid leads to "fluid memory", which propagates past "deformations" to the future [57, 58]. Fractal calculus has been shown to be a useful tool in the modeling of mechanical memory of viscoelastic materials [21]. Therefore, within bacteria, the viscoelasticity of the medium has been most frequently modeled with the fractional Langevin equation [13, 59, 21, 14, 60]. We should also note that the diffusion of molecules within homogeneous protein solutions has been successfully modeled using the fractional Langevin equation [61] and the MSD's of

the fractional Langevin motion are ergodic, Fig. 1.1C [21]. Finally, as with all previously mentioned models of sub-diffusion, fractional Langevin motion also results in anti-persistent behavior, in that when a molecule moves the medium "pushes back" [14]. The corresponding VAF $C_v^\delta(\tau)$ shows a consistent negative peak when δ and τ are equal for all measurable δ and τ (Fig. 1.1F) [13]. This behavior indicates that there is a "restoring force", causing the anti-persistent behavior over a large range of timescales due to the elastic nature of the medium. There are two major biological implications if a cell's cytoplasm is a viscoelastic medium; (1) molecules would take longer to reach distant targets than a freely diffusing molecule; and (2) molecules would retrace their previous locations, which could have interesting implications for the timescales of any process which depends upon two molecules coming together.

Continuous time random walk (sub-diffusion)

A fourth type of sub-diffusion behavior is described by the Continuous Time Random Walk (CTRW) model. In the CTRW the diffusion of a molecule is modelled as jumps on a lattice with random waiting times between individual jumps. The waiting time distribution follows a power law probability distribution, leading to large heterogeneities when comparing the MSDs of individual molecules. The proposed biological mechanism behind the CTRW are binding events along a trajectory [14], whose power law distribution for waiting times has been observed before [62]. The long tail of the waiting time distribution leads to the breaking of ergodicity, where the ensemble average does not equal the time averages of individual trajectories [15]. Here we should note that CTRW does not show anti-persistent behavior and the VAF does not have a

negative peak, Fig. 1.1D.

1.6 Diffusion in the cytoplasm

The cytoplasm is the largest compartment of a bacterial cell and the main reaction chamber for essential cellular processes such as signal transduction, protein degradation and gene regulation. As diffusion is the main means for bacterial macromolecules to reach their target sites in the cytoplasm, it is important to understand how the properties of bacterial cytoplasm influence diffusion.

A defining difference between the cytoplasm of eukaryotic cells versus bacteria cells is the level of crowding. For instance, in bacterial cells the concentration of proteins was measured at $200g/L$ in laboratory growth conditions, whereas that in mammalian cells was measured at 50-$100g/L$ [63, 64]. Under special conditions such as increased osmotic stress, the macromolecular concentration in bacterial cells can approach that of protein crystals [65]. The extreme crowding of cytoplasm has a massive influence on the diffusive properties of cytoplasmic molecules and is likely the main origin of sub-diffusion.

Diffusion of particles of different sizes

Early studies characterized the diffusion of fluorescent proteins (FPs) in live *E. coli* cells using FRAP [2]. By bleaching half of a cell and monitoring the fluorescence recovery, the diffusion coefficient (D) of GFP was determined at $\sim 8\mu m^2/s$. This D value is ~ 10 times slower than that in water [66] and ~ 4 times slower than that in the cytoplasm of eukaryotic cells [67], suggesting that the bacterial cytoplasm is indeed highly viscous, and that the diffusion of

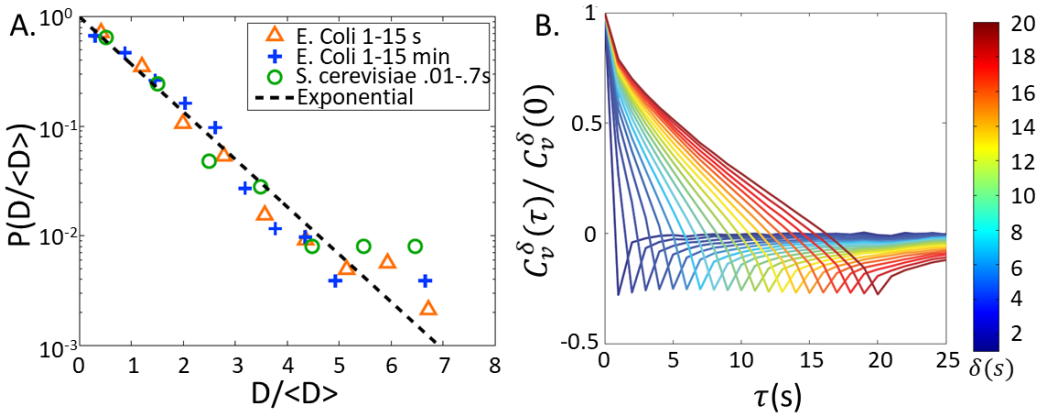

Figure 1.2: A. Dynamic Hetergenaity of individual mRNAs: the probablity density function of the diffusion coefficent of individual mRNA molecules normalized by their mean (Figure from [23]). B. The VAF of the mRNA resembles that of diffusion within a viscoelastic medium (Figure from [23]).

macromolecules in the cytoplasm could set the reaction timescales for certain cellular processes [2].

Interestingly, when a FP is fused to proteins of different molecular weights (MWs), although the trend holds true that the larger the MW is, the lower the diffusion coefficient, the quantitative relationship is different from what would be expect from the Stokes-Einstein equation [68]. Instead of scaling with MW with an exponent of $-1/3$, the experimentally measured scaling exponent is between -0.5 to -0.8 for proteins [65, 69]. As such, proteins exhibit a rapid reduction of D as MW increases. As we discuss below, other properties of proteins and the bacterial cytoplasm are likely responsible for this behavior.

For large and non-globular molecules such as mRNAs labeled with the MS2-FP fusion system [70] (MW $>$ 2 MDa), studies observed sub-diffusive motions with an MSD exponent α of $\sim$ 0.7 on timescales from seconds to

minutes [70, 14, 60, 41]. The sub-diffusive behavior did not appear to be dependent on the growth condition or the genetic backgrounds used in the experiments, as the exponent α remained similar under various conditions [14]. In one mRNA study two diffusive states were qualitatively observed, with one essentially immobile ("trapped") and the other freely diffusing throughout the cytoplasm. With these results, a model in which the heterogeneous, crowded cytoplasm traps/cages individual mRNA molecules was proposed [70]. A more recent study found that the mRNA molecules exhibited dynamic heterogeneity through time and space and was ergodic (Fig. 1.2A)[41]. Intriguingly, the diffusion coefficients of individual mRNA molecules followed an exponential distribution, showing more of a continuum than two distinct states. It should also be mentioned that a similar trend was also found for mRNA within Yeast cells, suggesting the behavior may be a universal trait, Fig. 1.2A. Notably, VAF analysis of these studies all showed anti-persistent behaviors over various timescales (Fig. 1.2B), suggesting that the mRNA's diffusive motion resembled that of fractional Langevin motion, i.e. mRNA molecules diffused within an viscoelastic medium [13, 23].

A recent study explored the diffusive properties of even larger particles in the bacterial cytoplasm. GFP-fused avian reovirus protein μNS, when expressed at different levels, self-assembles into large particles of different sizes. Parry et al., tracked the diffusion of these nanoparticles in *E. coli* and *C. crescentus* cells and found these large particles exhibited different diffusion properties when compared to molecules of smaller sizes.

First, the MSD curve of these nanoparticles showed sub-diffusive behavior

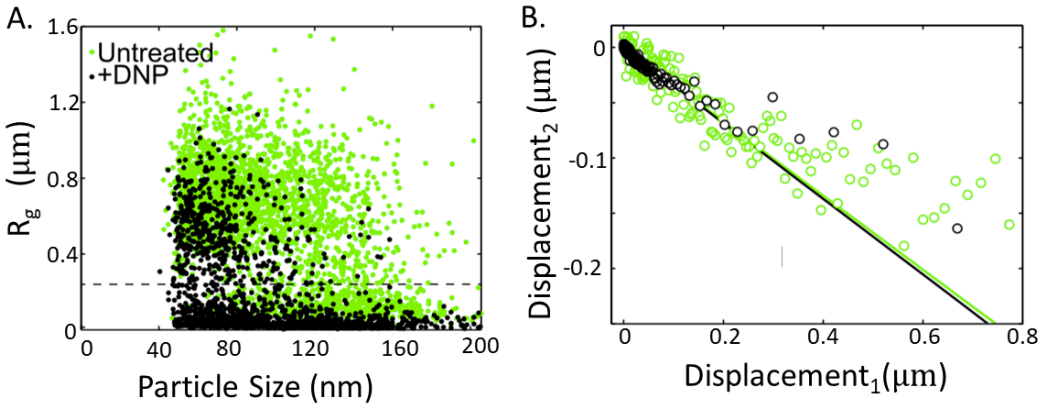

Figure 1.3: A. The radiation of gyration (R_G) of individual trajectories vs. the particle size for individual GFP-fused avian reovirus protein μNS particles, without (green) and with (black) ATP-depletion (DNP) [22]. B. The anti-persistent behavior of adjacent displacments for the same data in A. Here the directionallity was assigned a negative value if the second displacement was in the opposite direction of the first.

that was qualitatively similar to what was observed for mRNA [70]. They also exhibited two subpopulations, one immobile and one mobile. The presence of these two populations was independent of the corresponding particle size and the metabolic activity of cells, but the fractions of the two populations varied with both, Fig. 1.3A. The displacement distribution of nanoparticles was not Gaussian, and the larger the particles, the more they deviated from that expected for Brownian motion.

Second, the mobility of nanoparticles was related to the metabolic state of the cells: in metabolically inactive cells (ATP-depleted for example), there were more immobile particles (Fig. 1.3A), the MSD exhibited non-ergodic behavior, and the larger particles deviated from Brownian motion to an even greater extent when compared to smaller particles ($< 30 - 40nm$) [22]. These

results indicate that smaller particles within the cytoplasm "see" the cytoplasm as more of a fluid medium and the apparent diffusion coefficient of particles is greatly affected by the metabolism of the organism. Note that the differential diffusive behavior of small and large molecules/particles in the bacterial cytoplasm have also been reported under stressed conditions. In osmotically upshifted cells, the diffusion of GFP (quantified using FRAP) was found to decrease drastically compared to un-shifted cells [71, 72], while small molecules such as sugar molecules remained mobile and freely diffused throughout the cell [72].

Finally, these nanoparticles showed an anti-persistent behavior when the correlation of adjacent displacements was analyzed (Fig. 1.3B), suggesting that these particles have a preference to return to their previous positions [54]. As such, the bacterial cytoplasm was proposed to have glass-like properties, which affects the diffusion of molecules of different sizes differentially [22].

How does the bacterial cytoplasm behave like a glass-forming liquid? The differential mobility of small and large molecules/particles, the presence of the mobile and immobile states, non-ergodicity (in metabolically inactive cells) and the anti-persistent behavior, all suggest that the highly crowded cytoplasm likely traps particles in pockets/cages and that the cytoplasm is near its glass transition, at least in the metabolically inactive cells. These cages would confine the molecules/particles until the surrounding molecules in the cages rearrange themselves, likely by mechanic perturbations resulting from various enzymatic activities [22]. In metabolically inactive cells, the local cages would persist for a longer period of time compared to normal cells, explaining why

molecules are trapped in heterogenous pockets for longer times and why the deviation from typical Brownian motion grows larger. This effect directly links the timescales at which the cytoplasmic medium rearranges to the metabolism of the cell, providing a useful window to investigate bacterial cell metabolism. Notably, similar responses of chromosomal loci and an outer membrane protein (discussed later) to the cell's metabolism [60, 73] have also been reported, suggesting that it may be a universal rule that active metabolism of the cell increases the diffusion of molecules beyond what could be caused by simple thermal motion alone.

Many questions remain unanswered. Exactly how, at the molecular level, does the metabolic activity of a cell perturb the local cages in the cytoplasm? How are the diffusive dynamics of trapped molecules/particles influenced by the relative sizes of the surrounding molecules and chemical compositions? Do these glass-like properties influence any cellular processes in metabolically active cells, considering that most molecules in cells are likely too small to exhibit these effects with active enzymatic activity? Also, given that the cytoplasm of metabolically active cells exhibited only a fraction of the behaviors for a medium with "glass like properties", how should the viscoelastic properties, ergodicity, and specific distributions of dynamic diffusion coefficients that were seen for the mRNA molecule [13, 23], chromosome (discussed later), and nucleoid associated proteins (discussed later)[74] be incorporated into the theory? Finally, with this system a study should be done to investigate the timescales over which the μNS trajectories show anti-persistent behavior (the velocity autocorrelation function) to determine whether the behavior is con-

sistent with the other studies [13, 23, 74].

Diffusion of molecules of different surface properties

In the bacterial cytoplasm, the diffusive behavior has also been shown to be influenced by the surface properties of molecules. An early example came from the observation that the addition of a small (< 1 kD) but highly charged 6xHis tag to GFP caused a two-fold reduction of its D value in *E. coli* cells [2]. Another study systematically modified the surface net charge of GFP from -30 to 25 across multiple bacterial species and found that the most positively charged GFP variants had a D value of 100-fold slower than those of the negatively charged ones, likely caused by their electrostatic interactions with negatively charged ribosome, Fig. 1.4. Interestingly, the study pointed out that as the majority of cytoplasmic proteins in most bacteria are negatively charged and it is possible that these organisms evolved to limit nonspecific interactions with the ribosome in order to maintain a sufficient diffusion coefficient for its cytoplasmic contents [75].

A note of caution is that the D measurement of relatively small molecules such as GFP in above studies were done using FRAP. As mentioned earlier, FRAP is an ensemble method and unable to differentiate different types of diffusion and corresponding transition kinetics [2]. SMT of freely diffusing small protein molecules in live bacterial cells has been difficult in the past because of the molecules' relative fast diffusion. However, with recent development of bright organic fluorophores such as the Halo-JF dye system, and fast, sensitive

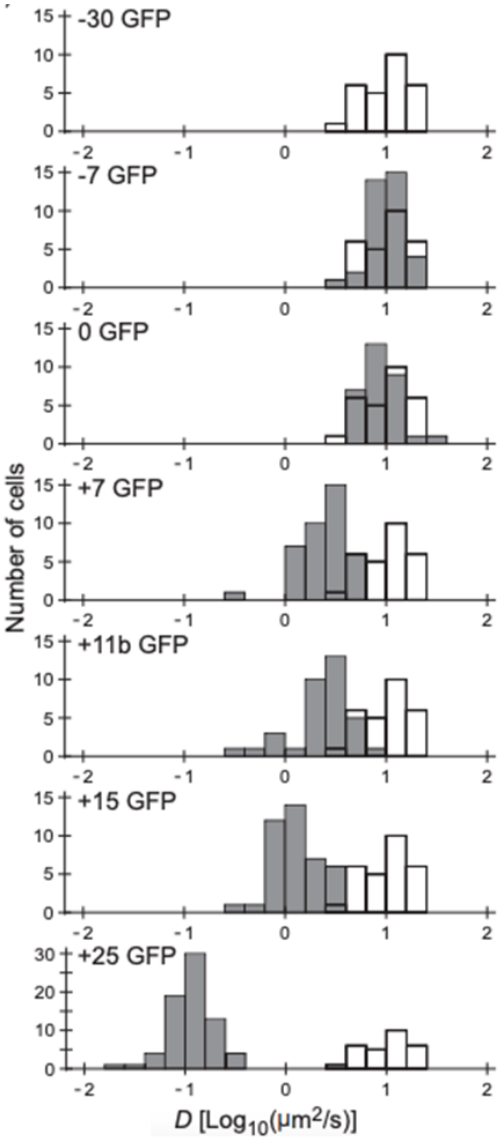

Figure 1.4: The relation between the charge of GFP and their diffusion coefficents: the filled histogram shows the distribution for the charged particle referenced in the individual subplots and the empty histogram shows the diffusion coefficents of the -30 GFP in each subplot for reference [75].

cameras such as the new generation of Scientific cMOS cameras, it is foreseeable that new information of the bacterial cytoplasm and dynamic interactions of normal-sized protein molecules with their interacting partners will emerge.

1.7 Diffusion in the nucleoid

The majority of the bacterial cytoplasm volume is occupied by the nucleoid, an enormous DNA-RNA-protein complex. The macromolecular structure and compaction of the nucleoid are maintained and regulated by small RNAs and many proteins such as histone-like nucleoid-associated proteins (NAPs) [76], topoisomerases [77] and the structural maintenance of chromosome (SMC) proteins [78]. The chromosome also dynamically rearranges when exposed to different stimuli [79]. Consequently, the organization and dynamics of the nucleoid itself influences how DNA binding proteins such as RNAP and transcription factors (TFs) find their targeting DNA sites.

For instance, chromosomal DNA loops [80] play important roles in transcription regulation and the overall compaction of the chromosome [81, 82]. DNA loops form when specific chromosomal regions come into contact with each other in space and the ends are restrained by protein binding. Chromosomal DNA segments of different genes could also be spatially positioned in proximity with each other to form the scaffold of the so-called transcription factories for RNAP and transcription factors binding. In *E. coli* and *B. subtilis* RNA polymerases were shown to form spatial clusters, where the synthesis of rRNA takes place [83]. Due to the local high concentration of RNAP, the diffusion of genes and/or transcription factors into and out of the RNAP clusters

is a likely mechanism of transcription regulation. Therefore, it is important to understand the diffusive properties and associated time scales of the dynamics of the chromosome and its interacting proteins [84].

Diffusion of chromosomal DNA

One important consideration of describing the diffusion of chromosomal DNAs is that chromosomal DNA is a polymer itself, thus its diffusive dynamics are different from that of any non-tethered particles within the cytoplasm. Note that while there were only limited numbers of studies on the chromosome's dynamics in bacteria, it has been shown that the general diffusive properties of the chromosome are conserved across different bacterial species.

In one study Weber et al. [14] used the ParB-GFP/parS system to label chromosomal loci in both *E. coli* and *C. crescentus* [85]. The labeled chromosomal loci exhibited sub-diffusive motion with an MSD exponent $\alpha \sim 0.4$. Under different perturbation conditions, although the D varied over $\sim$ 4-fold, the α value was unchanged, indicating that the dynamics of these individual loci are likely dominated by one universal physical process, Fig. 1.5A. The α value is also different from that of mRNA molecules measured using the MS2-GFP system ($\alpha \sim 0.7$) [70, 14, 23]. This difference suggests that the physical interactions of chromosomal DNA and mRNAs with their surroundings are likely very different. Incorporating the chromosomal polymer property into the diffusion model lead to an exponent of ~ 0.5 [57], indicating that additional factors must be at play. Interestingly, when the viscoelastic property of the cytoplasm (modeled by fractional Langevin motion [21]) was incorporated together with the polymer model, an exponent of $\alpha \sim 0.35$ was predicated,

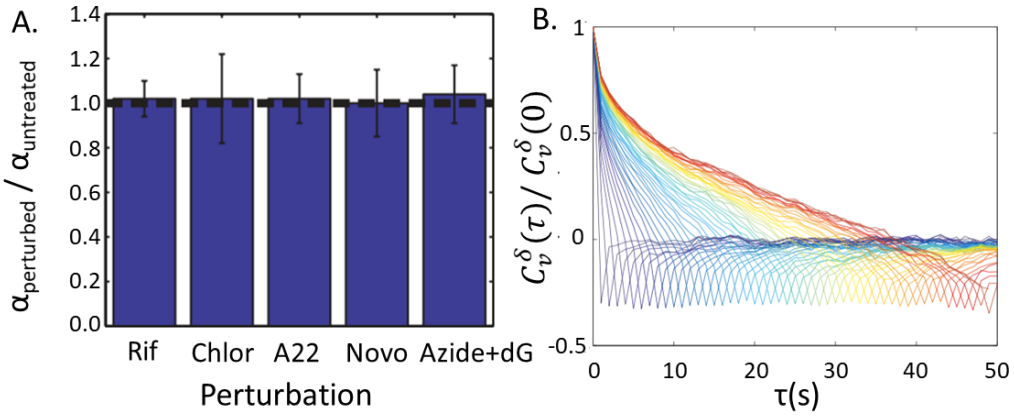

Figure 1.5: A. The behavior of the DNA's subdiffusive diffusion remains the same when exposed to different perterbations: the exponent of the MSD curve (α) remains the same when exposed to many different conditions (Figure from [14]). B. The VAF of the DNA resembles that of diffusion within a viscoelastic medium (Figure from [13]).

matching experimental measurements. Correspondingly, the velocity correlation function showed long timescale correlations for chromosomal loci, Fig. 1.5B. (and mRNAs too, Fig. 1.2B). These results hence provide strong support that the cytoplasm possesses viscoelastic properties that create "fluid memory" [13, 14].

One interesting discrepancy of this work is that a viscoelastic cytoplasm modelled by the fractional Langevin equation is fundamentally different from a cytoplasm with glass-like properties as what was proposed by Parry et al. A recent work by Sadoon et al., shed light on the discrepancy [74]. In this study the diffusive behavior of the DNA binding protein H-NS was investigated. The histone like H-NS oligomerizes on DNA and regulates the expression of $\sim 5\%$ of the *E. coli* genome. Using SMT of mEos3.2-fused H-NS, the apparent D value

was determined to be similar to that of the chromosomal loci with $\alpha \sim 0.6$, suggesting that the diffusion of H-NS is likely linked to that of the chromosome. The velocity autocorrelation function showed characteristics that were consistent with previous studies [13, 14, 41] suggesting a viscoelastic cytoplasm as modeled by the fractional Langevin equation [13]. Interestingly, when they quantified the complex modulus of the medium as a function of frequency (.1 sec^{-1}-20 sec^{-1}) the bacterial cytoplasm showed a glass-like transition over the different timescales, suggesting that the cytoplasm exhibited properties as reported by Parry et al [22].This study suggests that the cytoplasm of bacteria behaves differently at different timescales, highlighting the importance of taking timescales of different cellular processes into consideration.

In another study, Weber et al. found that their previous model viscoelastic cytoplasm coupled with the DNA polymer model was inadequate to capture the temperature dependence of the diffusion of labeled chromosomal DNA loci. The apparent diffusion coefficient D of chromosomal DNA loci scaled exponentially with temperature, termed "super-thermal", instead of linearly as predicted in the Stokes-Einstein equation when the system is at equilibrium [60]. Most interestingly, this super-thermal diffusion only existed in cells of active metabolism - in cells depleted of ATP, D scaled linearly with temperature as expected. These results indicate that the non-equilibrium state of the cell, most certainly caused by enzymatic activities, leads to "faster" diffusion than what would be produced solely by thermal fluctuations. Here the influence of metabolism on diffusion is consistent with the previously discussed study done by Parry et al [22].

Diffusion of DNA binding proteins

Since the chromosome is within the same compartment as the ribosomes within bacteria, all of the components that regulate the conformation of the chromosome, transcription and translation must function together within the same environment. Since the diffusion of chromosomal DNA is very small compared to that of DNA-binding proteins, different diffusive states of DNA-binding proteins, judged by their differential apparent diffusion coefficients, are commonly used to identify the bound and unbound states, providing an invaluable technique to study protein-DNA binding kinetic and functions in live cells.

An early SMT experiment done by Elf et al probed the binding of the transcription factor LacI to its specific chromosomal binding site lacO. While both 1d and 3d diffusion had been proposed as the mechanism for how transcription factors find their specific DNA targets in the presence of overwhelmingly nonspecific chromosomal DNA [86], the authors found that the a single LacI dimer spends the majority of its time (90%) performing 1d diffusion along the DNA, demonstrating this mechanism *in vivo*. A similar result was found for RNAP, which spends 85% of its time to bind non-specifically within the nucleoid [40]. In another recent study, the diffusion dynamics of Gyrase in *E. coli* was investigated using SMT [77]. Gyrase helps maintain the supercoiling state of the chromosome, which has a large effect on transcription [79, 81]. It was found that the average time gyrase molecules spent in the specific DNA bound states is $\sim$ 2 sec, with replication-proximal gyrase molecules having longer dwell times ($\sim$ 8 s). Such a difference suggests that different gyrase molecules may work at different capacities depending on the local topologi-

cal need, highlighting the unique power of SMT as an imaging technique to identify spatial information in live cells.

Along the same line, SMT studies helped resolve a discrepancy between biochemical and microscopic data regarding the spatial arrangement between transcription and translation in bacterial cells [87, 88, 89]. Biochemical studies showed that in bacterial cells translation occurs co-transcriptionally when mRNA is still being transcribed and physically attached to the DNA. Electron and fluorescence microscopy however showed that ribosomes are excluded from the nucleoid while RNAP is predominately nucleoid-associated [40, 89]. Using SMT of the ribosomal protein L1 and S2 tagged with mEos2, the diffusion coefficients of the free subunit and the incorporated, translating ribosome were found to be significantly different, and that the free subunits could diffuse freely throughout the nucleoid. Therefore, ribosome could assemble inside the nucleoid to initiate translation. Fully assembled, translating ribosomes, however, are mainly excluded from the nucleoid, suggesting that as translation is initiated, the transcribing mRNA could gradually move out of the nucleoid to continue translation. Indeed, such movement of actively transcribing gene loci has been observed in *E. coli* cells [90].

1.8 Diffusion in the cell envelope

The cell envelop of gram-negative bacteria has three layers, the outer membrane (OM), the inner membrane (IM), and the space in between where cell wall resides (periplasm). Gram-positive bacteria do not have the outer membrane but have a thick cell wall and an inner membrane. The outer mem-

brane acts as the first barrier between the cell and the environment for gram-negative bacteria. It is rich in β-barrel proteins, which allows small molecules to access the periplasm and cytoplasm through the inner membrane [91]. Another function of the outer membrane comes from its mechanical properties, as a recent study predicts that the outer membrane's β-barrel proteins play a large part in the ability of the cell to handle external forces [92]. The periplasm of gram-negative bacteria is often described as being "highly viscous" [93, 91, 94, 95, 96, 97] and contains a thin layer of peptidoglycan, or cell wall, although in reality it is likely not much more viscus than the cytoplasm [98, 99, 100]. The peptidoglycan layer dictates the cell shape and allows the cell to survive osmotic stress. The incorporation of this layer during division has been shown to be a major driving force for proper constriction [44]. Finally, the inner membrane directly links molecules in the cytoplasm to the environment on the outside and is important for a multitude of different signal transduction processes. The organization of bacterial membranes has been shown to be highly regulated and likely composed of many scattered microdomains [101, 102, 103].

Diffusion within the outer membrane

Despite the importance of the outer membrane and its associated outer membrane proteins (OMPs), (surprisingly) there are relatively few studies in which the diffusive behaviors of OMPs were investigated when compared to that of cytoplasmic proteins.

In an early work, where OMPs were nonspecifically labeled with a dye-conjugated reactive succinimidyl ester and chased with dye-free medium, it

was found that a significant proportion of the outer membrane proteins OMPs remained immobile in particular at the cell poles [104]. This observation is consistent with the notion that cell poles are essentially metabolically inert and stable.

For OMPs not specifically targeted to cell poles, many of them were found to be largely confined as well. SMT experiments of the outer membrane λ receptor protein tagged with large beads (20 to 500 nm) showed that that a subpopulation of the receptor was confined in small domains of 20 – 50 nm, and that a relatively faster population explored regions about 100-300 nm in size (Fig. 1.6A) [105, 106]. SMT of two other outer membrane proteins labeled with fluorescent antibodies, the porin OmpF and the cobalamin receptor BtuB, showed that OmpF was confined to domains of $\sim$100 nm in diameter, similar to the λ receptor, while BtuB was much more mobile with a D of 0.05 $\mu m^2/s$, an order of magnitude larger than that of OmpF [107].

What determines the differences in the mobility of OMPs and why do some OMPs exhibit confined diffusion? On one hand, depleting ATP, or inhibiting cell wall synthesis, caused a significant further reduction of the mobility of the λ receptor at long time scales [73], similar to what was observed for cytoplasmic proteins and the chromosome [22, 60]. On the other hand, because λ receptor is anchored to the cell wall covalently, it was proposed that the constant and dynamic energy-consuming reconstruction of the peptidoglycan layer underlies the diffusive behavior of the λ receptor [73]. A later study that quantified the dynamics of OmpA with and without the ability to bind the cell wall however showed essentially the same immobility, arguing against this hypothesis [109].

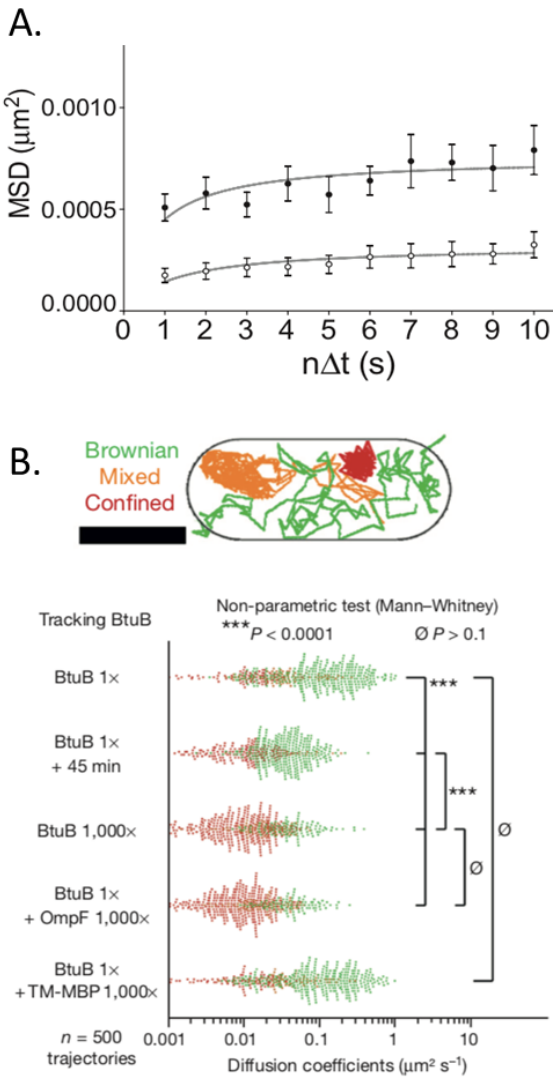

Figure 1.6: A. The confined diffusion of the OMP λ receptor with the filled circles calculated using the fast particles and the open circles the slow (Figure from [106]). B. Top shows an illustration of the colors representing the diffusive states of the individual molecules. Bottom shows how the diffusion of individual BtuB (OMP) was affected by the addition of different amounts of more BtuB or non-interacting OmpF. The addition of an engineered maltose binding protein with an single transmembrane helix (TM-MBP) was also used as a control (Figure from [108]).

Then the work of Rassam et al. showed that protein-protein interactions within the outer membrane appear to play an important role in restricting OMP diffusion. Using SMT, a mutated BtuB protein unable to interact with its cytoplasmic membrane protein partner TonB showed $>$ 10-fold increase of mobility. Interestingly, even nonspecific protein-protein interactions were shown to be important. Rassam et al. further showed that two other OMPs, Cir and BamA, which do not interact with BtuB directly, clustered with BtuB in 0.5-μm diameter "islands" on the outer membrane of *E. coli* cells. When the diffusion of BtuB was measured *in vitro* in a supported lipid bilayer made from *E. coli* membrane extract (Fig. 1.6B), SMT of BtuB showed Brownian diffusion at low concentrations. When the concentration of BtuB or a non-interacting OMP OmpF increased, BtuB exhibited orders of magnitude reduced diffusion and increased confinement. These results strongly suggest that the mechanism behind the previously observed confined diffusion was due to the "promiscuous" interactions among OMPs in confined areas of the outer membrane, which was proposed to be individual islands of different molecular compositions [108]. The proposed OMP islands formed by non-specific protein-protein interactions however need to be further verified. In particular, it would be interesting to examine whether all or just a few specific OMPs make up these islands, whether the characteristics of these OM islands vary with metabolism, and how the response changes the diffusion of other OMPs.

Diffusion within the periplasm

Surprisingly, only a few studies have investigated the diffusive behaviors of proteins in the periplasm. In an early study where the diffusion of the maltose-

binding protein (MBP) within the periplasm of *E. coli* was monitored using FRAP, the lateral diffusion coefficient of MBP was found to be at 0.009 $\mu m^2/s$ [110]. The extremely small diffusion coefficient was later shown to result from the harsh experimental conditions used to permeabilize the cells. Later FRAP studies found that the diffusion coefficients of FP tagged periplasmic proteins were $\sim 3\mu m^2/s$ [98, 99, 100], only slightly smaller than that in the cytoplasm. Additionally, when under osmotic stress (water leaves the cytoplasm and moves into the periplasm), the diffusion coefficient of these periplasmic proteins increased ~ 3 fold, similar to what was observed for the cytoplasm [100]. Most interestingly, the periplasms of multiple gram-negative bacteria have been shown to form heterogenous, diffusion-confined domains, suggesting that the proteins in the periplasm likely exhibit a level of crowding that influences each other's diffusion dynamics [98, 111, 112]. Clearly, further investigations especially with SMT methodologies are needed to elucidate the diffusion dynamics of proteins in the periplasm.

While no SMT studies have been done on a purely periplasmic protein at this time, various studies have quantified the diffusive properties of the enzymes responsible for maintaining the peptidoglycan layer during cell division (mostly inner membrane proteins). Of particular interest to the study of diffusive behavior within bacteria, many of these proteins show super-diffusion, whose corresponding velocities are likely directly linked to their state. Furthermore, studies are beginning to show how the information within the dynamic behavior of molecular assemblies within the cytoplasm are propagated into the periplasmic compartment.

In most bacteria, for cell division to take place, a large macromolecular complex, the divisome, must form and direct the synthesis of septal peptidoglycan. The formation of this complex is initiated by FtsZ, a tubulin homolog, which polymerizes at the middle of dividing cells. While FtsZ SMT studies have shown that the individual monomers of the FtsZ filaments are stationary [113], recent works utilizing total internal reflection fluorescence microscopy have shown that the filaments themselves show directional movement, the result of treadmilling [44]. The FtsZ filaments' dynamics are thought to direct/coordinate the incorporation of the peptidoglycan and organize many of the other proteins in the divisome.

Interestingly, even through many of the enzymes important for septal peptidoglycan incorporation show the same super-diffusive motion, the mechanisms behind their motions seem to vary between different bacterial species. In bacillus subtilis it was found that the velocity of bPBP2b (penicillin-binding protein) was directly linked with the velocity of the FtsZ filaments and that the velocities of these components were directly linked to septum closure [46]. Similarly, in *E. coli* it was shown that the velocity of the synthase enzyme bPBP3 (FtsI) was also directly correlated with the velocity of the FtsZ filaments, but interestingly the velocity of the two were not limiting in terms of septum closure [44]. Lastly, unlike the other two species in Streptococcus pneumoniae it was recently found that the bPBP2x:FtsW complex showed directional motion but its velocity was independent of the velocity of the FtsZ filaments [45]. Considering the similarities in the diffusion dynamics and the rarity of directional motion within bacteria future work quantifying the mech-

anisms responsible for the diffusion of these enzymes is an exciting direction of study.

Diffusion within the inner membrane

Compared to proteins in the outer membrane, inner membrane proteins (IMPs) appear to be more mobile. The first SMT study of an IMP tracked the membrane-bound histidine kinase PleC fused with a YFP in *C. crescentus* cells. PleC localizes to the cell pole of Caulobacter cells and was shown to be important for the asymmetric cell division [114]. A subpopulation of PleC-YFP indeed was found at the cell pole and was largely immobile, and the other subpopulation diffused within the cell body with normal Brownian motion with a D of $\sim 0.01\mu m^2/s$. This observation suggests that at least some IMPs can freely diffuse throughout the entire inner membrane within *C. crescentus* cells [115]. Another IMP, TatA, forms large complexes ($\sim$ 600 KDa) with itself and the other two proteins TaB and TacC in the twin-arginine translocon [116]. TatA diffused faster than PleC with an apparent D of $\sim .13\mu m^2/s$ measured by FRAP [99]. Such a high mobility is comparable to what was observed within Eukaryotic membranes [117]. SMT of TatA-YFP in another study showed similar Brownian motion with a comparable D value, although the trajectories were not long enough to identify whether there were other diffusive modes at long time scales [118].

The relationship between the size of an IMP and its diffusion coefficient was also investigated. For TatA-YFP, for example, SMT found that the apparent diffusion coefficient value decreased when the number of TatA-YFP molecules in self-assembled complexes increased from $\sim$ 10 to 100. Interestingly, the

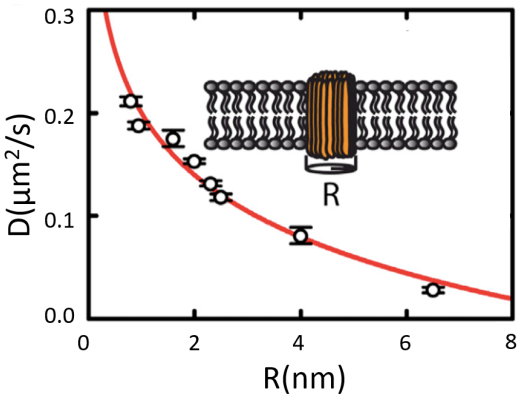

Figure 1.7: Diffusion coefficents of IMPs vs. the radius of the IMP (R) (Figure from [119]).

relationship can be strictly described as logarithmic, mimicking what has been observed for the relationship between the size of cytoplasmic proteins and the corresponding diffusion coefficient [118, 69]. Another work by Oswald et al investigated the diffusion coefficients of eight different inner membrane proteins and showed a clear relationship between the apparent diffusion coefficient and the radii, but not the molecular weight, of each protein (Fig. 1.7) [119]. This interesting finding suggested that it is the "amount/volume of the protein" that interacts with the membrane that dictates the apparent diffusion coefficient [120]. A more recent and exhaustive study has verified this finding, which investigated ∼200 membrane proteins within Bacillus subtilis [121].

Not all IMPs exhibit Brownian motion. The cytochrome bd-I complex (CydB), when fused to GFP, was found to form clusters of ∼ 100nm in diameter and contain approximately 76 cydB-GFP proteins per cluster. A similar logarithmic relationship between the apparent diffusion coefficient and the number of CydB within the complex as that of TatA-YFP was also observed.

However, some CydB-GFP clusters clearly exhibited confined diffusion. By fitting the MSD curves to a mobility-confinement model, the confinement zone of CydB-GFP, defined as an area in which the molecules can diffuse freely and above which a barrier confines the diffusion of the molecules was determined to have a diameter of 160nm [122]. A similar confinement zone was observed for another IMP, Tsr, a chemotaxis proteins that forms clusters at cell poles [123]. A few other IMPs that do not form clusters also showed sub-diffusive behavior based on their MSD plots, suggesting that sub-diffusion may be a common feature of IMPs.

What contributes to the confinement of IMPs? The inner membranes of *B. subtilis* and *E. coli* stained with membrane dyes Nile Red and Dil-C12 respectively, showed a clustered, heterogenous fluorescence distribution [119]. As it is known that both dyes are more specific for fluid rather than rigid regions of the membrane, it is likely that there existed fluid macrodomains on the membrane, and that these macrodomains may be responsible for the confined movement of IMPs. In eukaryotic cells it is known that cytoskeleton proteins such as actin is involved in the formation of membrane microdomains [124]. Therefore, it is not surprising that when the polymerization of the bacterial actin homolog MreB was inhibited, the apparent Ds of many IMPs increased, confinement disappeared, and that the proportion of Tsr molecules that experienced confined diffusion diminished significantly [119].

Based on these results, a general model behind the confinement caused by MreB was proposed [119], similar to what was proposed for actin in eukaryotic cells [125]. Filaments formed by MreB and its membrane anchors may act

as diffusion barriers/fences for inner membrane proteins , leading to their apparent confinement observed in the MSD plots. However, Lucena et al., found that the diffusion of > 200 IMPs along the long axis of the cell and the short axis did not exhibit any markable differences, arguing that the MreB filaments, which mainly form along the short axis of the cell, may not be confining the diffusion of the proteins using the same mechanism as actin, or that MreB filaments may not be as dense or well organized as long actin filaments in eukaryotic cells [121].

1.9 Summary

Within this chapter the various methodologies used to characterize diffusion within bacteria with a primary focus on single molecule tracking were described. We have described the different forms of analysis one can use to determine the behavior of diffusion as well as described the most common models of diffusion. We then went into detail, describing the studies that have characterized the diffusion within the different compartments of bacteria, and several different themes emerged from the various studies. First, in every compartment of the cell, diffusion is much more complicated than normal Brownian motion, with a vast array of different mechanisms leading to subdiffusion. Second, the metabolism of the cell can have a large impact on the diffusion of particles in any compartment of the cell, where the higher the metabolism the faster the diffusion of the particles. Third, the cytoplasm has viscoelastic properties and influences the diffusion of particles on a variety of different timescales. Four, diffusion within the cell envelope has not been

quantified as in depth as within the other compartments. And five, different protein-protein interactions can lead to different types of diffusion, including confinement. Finally, we would like to emphasize that no real consensus has been made for any compartment and no model can yet explain all the experimental work at this time, leaving room for many new discoveries.

Chapter 2

Reduction of Confinement Error in Single-Molecule Tracking in Live Bacterial Cells Using SPICER

[1]

2.1 Background

Single molecule tracking (SMT) is a powerful technique to probe possible functional states of biomolecules in living cells [126, 127, 42]. In a typical SMT experiment, a molecule's cellular positions are recorded by acquiring its fluorescent images consecutively at defined time intervals. From these images, a SMT trajectory, a time series of corresponding spatial coordinates of the molecule in reference to the cell, is extracted. From the statistical analysis of these SMT trajectories, different diffusive states of the molecule, each charac-

[1]Bohrer CH, Bettridge K, Xiao J. Reduction of confinement error in single-molecule tracking in live bacterial cells using SPICER. Biophysical journal. 2017 Feb 28;112(4):568-74.

terized by a different diffusion coefficient D, can be obtained. These diffusive states and the associated population percentages can provide valuable information regarding possible functional states of the molecule. Recent SMT in bacterial cells have indeed shed light into the working mechanisms of transcription factors [128, 129] , RNA polymerase [40, 51, 89], DNA polymerase [130], ribosomes [131], cytoskeletal proteins [113, 49] and more [132, 133].

In addition to measuring a molecule's diffusion coefficients, SMT experiments offer another significant advantage, which is to obtain transition probabilities of molecules between different diffusive states. These transition probabilities provide crucial information regarding the kinetics of state-switching such as the binding and unbinding rates of a protein molecule to its target site, and the lifetime of a particular functional state of the molecule [8, 10]. Such information is often difficult to obtain in live cells by other means.

Various algorithms based on the statistical analyses of a large number of SMT trajectories have been developed to obtain the transition probabilities and associated diffusive states from SMT experiments. Among them, the vbSPT algorithm developed by Persson et al. has proven robust [10]. vbSPT assumes a hidden Markov model (HMM) in which diffusing molecules make a memoryless jump in states defined by different diffusion coefficients, and uses a variational Bayesian approach to identify individual states and their associated kinetics [10].

Successful application of analysis methods like vbSPT requires the correct identification of different diffusive states, which are characterized by unique diffusion coefficients. However, in bacterial cells, the small cell size (1-2μm) spatially confines a molecule's diffusion, such that the measured apparent diffusion coefficient D_{app} of a molecule appears smaller than the actual value, leading to difficulties in identifying the correct diffusive state. A common practice to minimize confinement effects is to use displacements measured only along the long axis of the cell, where molecules would experience the least confinement; we refer to this method as "1d" analysis throughout the work [134, 51, 10]. However, using information only along one dimension leads to a less accurate determination of diffusion parameters than using all available dimensions. The reduced amount of data limits the available number of well-defined trajectories, which is particularly important for calculating transition probabilities [135, 136].

Here we present a new, simple algorithm, termed SPICER for Single Particle-tracking Improvement with Confinement Error Reduction. SPICER maintains the full length of a SMT trajectory and maximizes the amount of data used by calculating displacements in all dimensions available (2d or 3d) and only selectively switches to 1d (along the cell's long axis) when a molecule is likely to experience confinement. As such, the accuracy in determining both the diffusion states and transition probabilities is dramatically improved. We demonstrate the use of SPICER on both simulated and experimental SMT data of *E. coli* RNA polymerase (RNAP). The simple implementation of the SPICER

algorithm in SMT analysis should allow its wide application in probing *in vivo* dynamics of molecular events in small bacterial cells.

2.2 Operational principle of SPICER

To illustrate the operational principle of SPICER, we show in Figure 2.1 a schematic $3d$ SMT trajectory of a diffusing molecule in a typical rod-shaped bacterial cell with two cross sections along the long x, top panel, and short y, bottom panel, axes of the cell. The parameter R defines the confinement zone (red), and is the distance from the membrane boundary of the cell to the edge of the midcell region where the molecule diffuses freely and does not experience confinement (green).

In previous studies, to avoid confinement, only displacements along the cell's long axis were used for HMM analysis. We refer to this method as "$1d$" analysis in this work [134, 51, 10]. Consequently, in the $1d$ analysis the available data in the trajectory ω shown in Figure 2.1 , represented by a series of single step displacements, ω=$(\Delta r_1^{3d}, \Delta r_2^{3d}, \Delta r_3^{3d}, \Delta r_4^{3d}, \Delta r_5^{3d}, r_6^{3d})$ and containing information in all three dimensions ($3d$), $(r_j^{3d})^2 = \Delta x_j^2 + \Delta y_j^2 + \Delta z_j^2$, is reduced to a third of the original amount of data, $(r_j^{1d})^2 = \Delta x_j^2$. SPICER increases the amount of data available and reduces the proportion of data that experiences confinement by analyzing a modified trajectory, $w' = (\Delta r_1^{1d}, \Delta r_2^{3d}, \Delta r_3^{1d}, \Delta r_4^{3d}, \Delta r_5^{3d}, \Delta r_6^{1d})$, where displacements $\Delta r_{1,3,6}^2$ are calcu-

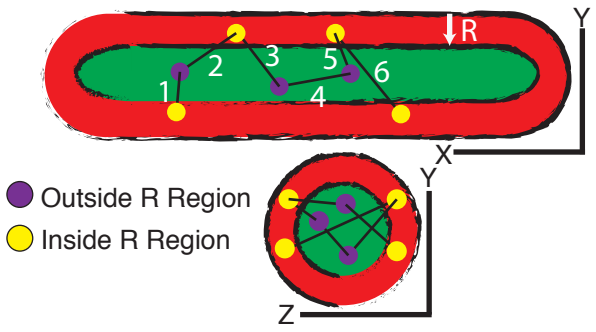

Figure 2.1: An example $3d$ SMT trajectory of a molecule in a rod-shaped bacterial cell. The purple filled circles are localizations of the molecule inside the confinement-free region (green). Displacements using these localizations as initial positions, such as displacements 2, 4, and 5, are calculated using their full $3d$ coordinates. Yellow filled circles are localizations of the molecule inside the R-region where the molecule experiences confinement (red). Displacements using these localizations as initial positions such as displacements 1, 3, and 6 are calculated using only $1d$ coordinates along the x (long) axis of the cell.

lated in $1d$ along the long axis of the cell to avoid confinement in the R region, while $\Delta r^2_{2,4,5}$, are calculated in $3d$, as the initial positions of these displacements are in the midcell and outside of the R region. As such, the full length of the trajectory is maintained, there is an increase of data being utilized with coordinates of all available dimensions, and a decrease in the proportion of data experiencing confinement. Note here that the same principle can be applied to $2d$ tracking experiments because of the symmetry of rod-shaped bacterial cells along the short axis.

An example of a $2d$ SMT trajectory modified by SPICER is shown in Figure 2.2, with the confinement zone shown in red and the freely diffusing region in green. Intuitively, the operational principle of SPICER is still justified for

$2d$ tracking data as displacements in the center of the cell (green) will have a higher probability to belong to true localizations outside the confined R-region. This is because a rod-shaped bacterial cell is isotropic along the short axis of the cell. By having a large number of trajectories sampling all possible positions, localizations in the periphery and center of the cell will still have high probabilities to be correctly identified as inside or outside of the R-region.

It is important to note that while applicable to $2d$ SMT, the use of SPICER on $2d$ tracking data is at a disadvantage when compared to $3d$ tracking, due to the lack of information along the third dimension. The uncertainty in the third dimension creates a chance that a small percentage of the displacements selected by SPICER as having no confinement will possess some confinement error, as indicated by the circled spot in Figure 2.2. Hence, the application of SPICER to $2d$ data results in a less significant improvement in the calculation of the different parameters when compared to the $3d$ tracking data.

Next, we demonstrate that switching dimensions within an SMT trajectory as described above does not modify the ability of SPICER to identify a set of most suitable parameters (diffusion coefficients D and transition probabilities P) describing the trajectory using the Maximum likelihood method [8].

The likelihood of having a diffusion coefficient D given a single displacement in d dimensions, $L(D|\Delta r_j)$, is proportional to the probability of having that displacement given the diffusion coefficient, $P(\Delta r_j|D)$, and is defined by

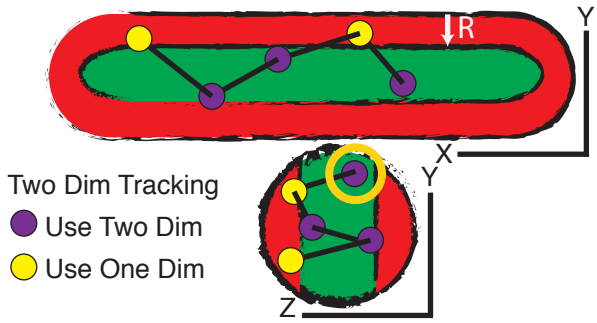

Figure 2.2: An example $2d$ SMT trajectory of a molecule in a bacterial cell. The purple circles are localizations inside the confinement-free region (green), and displacements calculated using these localizations as initial positions utilize their full $2d$ coordinates. Yellow circles are localizations inside the R-region and experience confinement (red). Displacements calculated using these localizations as initial positions only utilize coordinates along the x (long) axis of the cell. Both purple and yellow localizations are $2d$ projections of molecule positions in $3d$, and hence it is possible that a localization that appears to be outside the R-region is actually inside the R-region and experiences confinement (yellow hollow circle), but its full coordinates are used.

the following equation:

$$L(D|\Delta r_j) \propto P(\Delta r_j|D) = \frac{e^{-\frac{\Delta r_j^2}{4D\tau}}}{(4\pi D\tau)^{d/2}}, \tag{2.1}$$

where $\Delta r_j^2 = \Delta x_j^2$ for d=1, $\Delta r_j^2 = \Delta x_j^2 + \Delta y_j^2$ for d=2, and $\Delta r_j^2 = \Delta x_j^2 + \Delta y_j^2 + \Delta z_j^2$ for $d = 3$; D is the corresponding diffusion coefficient and τ is the time interval for each displacement. Equation 2.1 is the direct result of solving the diffusion equation with no barriers. If a molecule stays in one state as defined by a single diffusion coefficient D, the likelihood of having a particular trajectory specified by a series of experimentally measured displacements, w,

will be:

$$L(D|w) = \frac{e^{-\frac{\Delta r_1^2}{4D\tau}}}{(4\pi D\tau)^{d/2}} \times \frac{e^{-\frac{\Delta r_2^2}{4D\tau}}}{(4\pi D\tau)^{d/2}} \times ... \times \frac{e^{-\frac{\Delta r_j^2}{4D\tau}}}{(4\pi D\tau)^{d/2}}. \tag{2.2}$$

Maximizing the likelihood, L, with respect to D results in the well-known relation $\langle r^2 \rangle = 2D\tau$ for d=1, $\langle r^2 \rangle = 4D\tau$ for d=2 and so on [8]. In previous studies, the value d is set constant for all displacements in a trajectory. However, note here that the true diffusion coefficient D is independent of the d value used, that the probability of each displacement is independent of the other displacements at each time point, and that the likelihood L is an arbitrary multiplicative constant with no significance in its absolute value in isolation. Therefore, varying d values along a trajectory does not prevent maximizing the likelihood to find the best-fit parameter D.

To illustrate that d can be varied throughout a trajectory, we assume that a molecule has a trajectory w of N displacements, and spends v displacements within the R region and k displacements outside of the R region, with $v + k = N$. In the simplest scenario where the molecule exists only in one state, the likelihood of the molecule having a D value given the trajectory is:

$$L(D|w, R) = \frac{1}{((4\pi D\tau)^{1/2})^v} e^{\sum_i^v -\frac{\Delta x_i^2}{4D\tau}} \times \frac{1}{((4\pi D\tau)^{2/2})^k} e^{\sum_j^k -\frac{\Delta x_j^2 + \Delta y_j^2}{4D\tau}} \tag{2.3}$$

The log of the likelihood can be expressed as:

$$l(D|w, R) = -(k + \frac{v}{2}) \log(D4\pi\tau) - \sum_i^v \frac{\Delta x_i^2}{4D\tau} - \sum_j^k \frac{\Delta x_j^2 + \Delta y_j^2}{4D\tau} \tag{2.4}$$

Simplifying by substituting with v = N-k results in:

$$l(D|w,R) = -(\frac{k+N}{2})\log(D4\pi\tau) - \sum_{i}^{N}\frac{\Delta x_i^2}{4D\tau} - \sum_{j}^{k}\frac{\Delta y_j^2}{4D\tau} \tag{2.5}$$

Maximize the log of the likelihood L with respect to D by taking the derivative results in:

$$D = \frac{\sum_i^N \Delta x_i^2 + \sum_j^k \Delta y_j^2}{2\tau(N+k)} \tag{2.6}$$

Which can be further converted to the mean squared displacement of each dimension by

$$D = \frac{\langle \Delta x^2 \rangle + \langle \Delta y^2 \rangle \times k/N}{2\tau(1+k/N)} \tag{2.7}$$

Equation 2.7 holds true irrespective of the value of k, be k=0 or N. For $0 < k < N$, the diffusion coefficient D that best fits the system is the proportioned combination of the two mean squared displacements. Equation 2.7 further emphasizes that changing the value of d in a trajectory has no effect on the parameters obtained by maximizing the likelihood as long as the R-value, and hence the number of v or k displacements, is kept constant.

Equation 2.7 demonstrates that including a proportion of non-confined displacements along the short axis in the likelihood calculation will increase the calculated diffusion coefficient if there is confinement experienced by $\langle \Delta x^2 \rangle$. This is why SPICER is able to outperform even the $1d$ analysis. For example if there was no confinement $\langle \Delta x^2 \rangle$ and $\langle \Delta y^2 \rangle$ would be equal, but because the $1d$ analysis still experiences confinement in the cell poles, $\langle \Delta x^2 \rangle$ is smaller

than expected. By including data along the short axis with no confinement, outside the R region ($\langle \Delta y^2 \rangle > \langle \Delta x^2 \rangle$), the calculated diffusion coefficient rises when the likelihood is maximized, see Eq. 2.7.

2.3 Selection of an optimal R value

Before one can analyze SMT data with SPICER, an optimal R-value for a given experimental system must be identified. The R-value defines the size of the confinement zone, within which the displacements of a trajectory are calculated using only $1d$ coordinates along the long axis. Displacements outside of the R region, toward the center of the cell, are computed using the full coordinates available in $2d$ or $3d$, depending on the experimental setup.

Intuitively, the size of the confinement zone, or the R-value, is primarily dependent on how fast the molecule diffuses. Molecules diffusing quickly require a large R-value to avoid confinement, whereas molecules diffusing slowly do not. Therefore, for a mixed population of molecules, the optimal R-value, R_{opt}, should be set for molecules that diffuse the fastest. Consequently, it follows that at a given imaging speed, one can construct a lookup table so that each estimated D_{max} value of a system corresponds to an optimal R-value.

To create the lookup table, we simulated five sets of $3d$ SMT experiments in a rod-shaped cell with radius $r = 500$ nm and length $l = 2$ μm; each set contains 10,000 single-state SMT trajectories with a fixed D_{true} ranging from

0.4 to 4 $\mu m^2/s$, tracked with an imaging speed of 200 f/s. For each dataset, we varied the R-value systematically from 50 to 500 nm with 50-nm interval and used SPICER to identify the corresponding apparent D value (D_{app}) at each R-value. We then plotted the approximation percentage (D_{app}/D_{true}) of the dataset at different R values (Figure 2.3 A). The optimal R-value was then identified as the one at which the approximation percentage reaches a maximum. The reasoning is that if an R-value is correctly selected, D_{app} would contain the least confinement error in SPICER, and hence approaching maximally the D_{true} value.

As shown in Figure 2.3A, datasets with small diffusion coefficients reach their maximum apparent diffusion coefficients D_{app} at small R values, consistent with the notion that slowly diffusing molecules experience less confinement and hence the R-region would be small. However, for the dataset that has a $D = 4\ \mu m^2/s$, $R_{opt} = 450$ nm, indicating that for molecules diffusing faster than 4 $\mu m^2/s$, the small size of a bacterial cell itself confines diffusion regardless of how far away from the membrane the molecule is. In Figure 2.3B we plotted the optimal R-value for each simulated D_{true} value. It can be seen that R_{opt} monotonically increases with D. Note here that while this look-up table is coarse-grained with R-value changing in 50-nm and D-value changing in $\sim 1\mu m^2/s$ increments, finer grains on the order of 5-nm and 0.1 $\mu m^2/s$ are not necessary. The typical spatial resolution in an SMT experiment in live bacterial cells is in the range of $30 - 50$nm and that a change of 0.1 $\mu m^2/s$ in D does not lead to a significant change in the corresponding

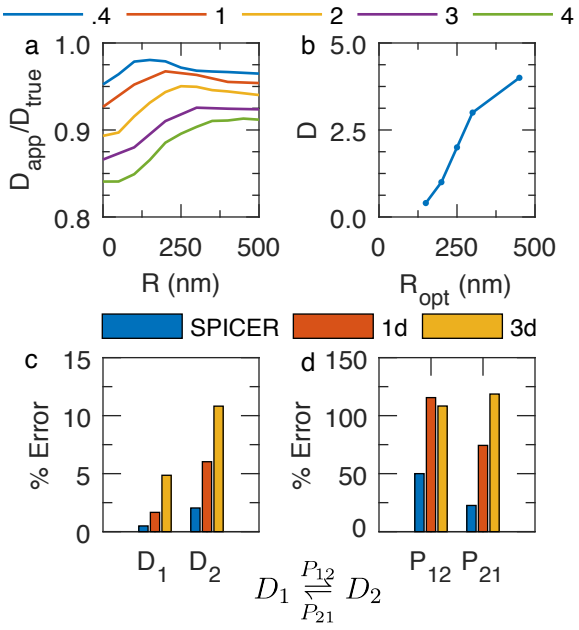

Figure 2.3: (a and b): Construction of a look-up table for finding optimal R-value in a $3d$ tracking system. (a) Approximation percentage (D_{app}/D_{true}) of five simulated systems at different R-values with D_{true} varying from 0.4 to $4\mu m^2/s$. (b) Optimal R-values at different diffusion coefficients are identified from (a) as the R-value at which the maximal D_{app}/D_{true} is reached. (c and d): Comparison of the performance of SPICER and conventional $1d$ and $3d$ analyses in identifying the diffusion coefficients (c) and transition probabilities (d) in a two-state system with $D_1 = 1\ \mu m^2/s$, $D_2 = .7\ \mu m^2/s$, and $P_{12} = P_{21} = .0244$. The percentage error is defined as $\frac{|X - X_{true}|}{X_{true}} \times 100$.

R-value within the 50 nm increment. Therefore, to use this lookup table, one can first estimate the largest diffusion coefficient of a given system using the $1d$ analysis, which approximates the true D value as it eliminates confinement error along the cell long axis, and then use Figure 2.3B to estimate the R_{opt} value. A similar simulation and lookup table using $2d$ SMT data is shown in Figure 2.4. A particular note here is that the look-up table is also related to the imaging speed, for that a fast diffusing molecule imaged with a slow speed (long time intervals between subsequent acquisitions) will naturally require a larger R-value to accommodate the longer distance it travels during the time. Therefore, it is important to construct the look-up table based on the actual imaging condition, as we described above.

Next, we verified whether the utilization of an optimal R-value in SPICER indeed improves SMT analysis when molecules exist in two different diffusive states. We simulated 25,000 trajectories of a two-state system. The two diffusion coefficients are $D_1 = 1\ \mu m^2/s$ and $D_2 = .7\ \mu m^2/s$, and the transition probabilities $P_{12} = P_{21} = .0244$, which corresponds to a transition rate of $\sim$ 5 per second. We then analyzed this dataset using $1d$ (only using displacements along cell long axis), $3d$ (using all displacements in three dimensions), or SPICER, in which the R_{opt} was chosen at 200nm (the larger D_1 at $1\ \mu m^2/s$) according to Figure 2.3B.

In Figure 2.3 C and D we plotted the percent error in each of the four parameters analyzed using the three methods. Clearly, $3d$ analysis led to the

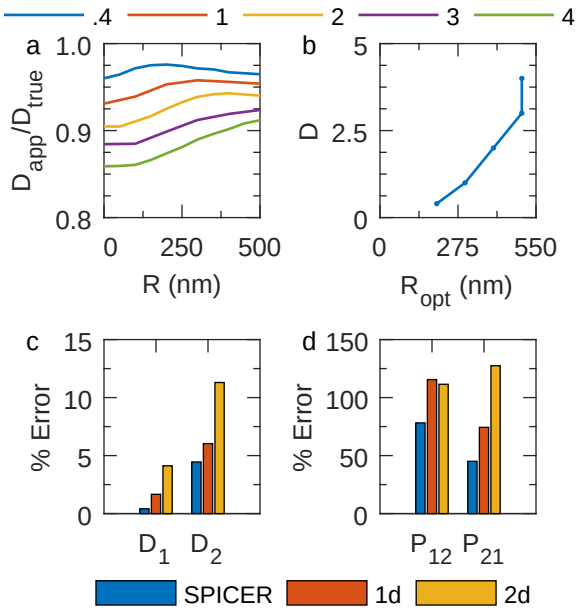

Figure 2.4: (a and b):Finding optimal R-values for $2d$ tracking systems (a) Approximation percentage (D_{app}/D_{true}) of five simulated systems at different R-values with D_{true} varying from 0.4 to $4\mu m^2/s$, tracking at an imaging speed of 200 f/s. (b) Optimal R-value lookup identified at different diffusion coefficients from (a). (c and d): Comparison of the performance of SPICER and conventional $1d$ and $2d$ analyses in identifying the diffusion coefficients (c) and transition probabilities (d) in a two-state system with $D_1 = 1\mu m^2/s$, $D_2 = .7\mu m^2/s$, and $P_{12} = P_{21} = .0244$. The percentage error is defined as $\frac{|X - X_{true}|}{X_{true}} \times 100$.

D_1	D_2	P_{12}	P_{21}	# of Traj
1	.4	.0476 (k=10/sec)	.0476	35000
1	.5	.0476 (k=10/sec)	.0476	35000
1	.6	.0696 (k=14/sec)	.0696	35000
1	.7	.0242 (k=5/sec)	.0387 (k=8/sec)	35000
1	.8	.0929 (k=20/sec)	.0464	35000
1	.9	.0714 (k=15/sec)	.0340	35000

Table 2.1: The parameters of the two state systems for the two SI figures S3 and S4.

highest amount of error for all the four parameters, consistent with the presence of significant confinement errors when displacements of all $3d$ were used under this condition. The $1d$ analysis showed improvement especially in the identification of D compared to the $3d$ analysis, but was significantly outperformed by SPICER, in which the percentage errors in all four parameters were the smallest. Note that localizations in the R-region of cell poles are still confined in SPICER, even though only their displacements along cell's long axis are used, same as that in $1d$ analysis. Nevertheless, SPICER outperforms the $1d$ analysis because in SPICER, the full coordinates of localizations outside of the R-region are given more weight than their corresponding $1d$ coordinates in the search for optimal parameters. We further verified that the same trend holds for a variety of systems with different diffusive parameters (Table 2.1, $3d$ Figure 2.5, $2d$ 2.6). These results demonstrate that by increasing the proportion of data containing full coordinates, SPICER with an optimal R-value identified from the look-up table (Figure 2.3B) indeed improves the accuracy in determining both diffusion coefficients and transition probabilities of a diffusive system.

2.4 SPICER improves accuracy in identifying states with close diffusion coefficients

One important criterion used by the HMM to identify different diffusive states is the difference between diffusion coefficients associated with each state. If the

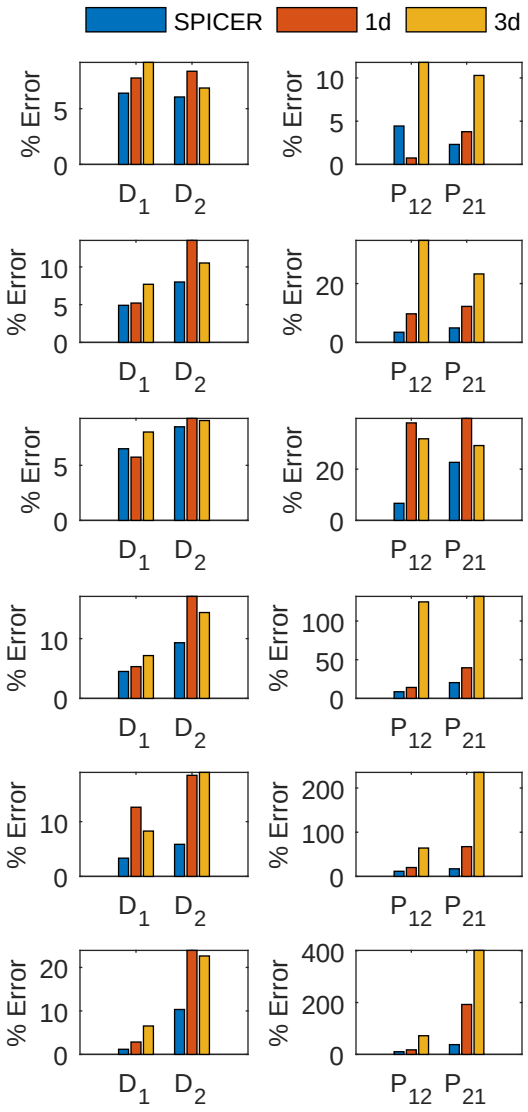

Figure 2.5: Percent errors in D_1, D_2 (left column) and P_{12}, P_{21} (right column) identified using SPICER, $1d$ and $3d$ analyses for different $3d$-tracking systems listed in Table S1. Each row in the figure corresponds to the same row in Table S1. In all the systems tested, SPICER outperforms the $1d$ and $3d$ analyses.

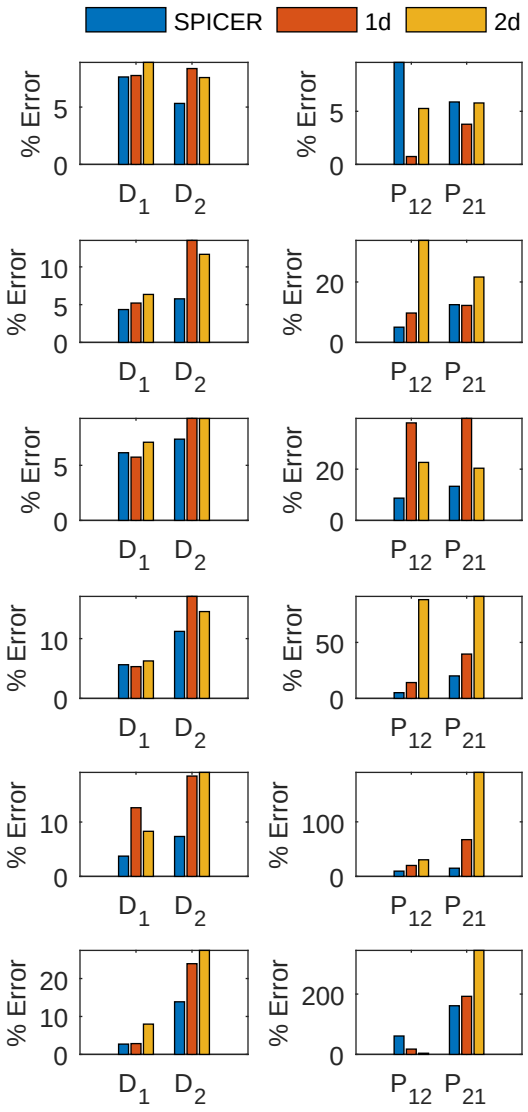

Figure 2.6: Percent errors in D_1, D_2 (left column) and P_{12}, P_{21} (right column) identified using SPICER, $1d$ and $2d$ analyses for different 2d-tracking systems listed in Table S1. Each row in the figure corresponds to the same row in Table S1. In all the systems tested, SPICER outperforms the $1d$ and $2d$ analyses.

diffusion coefficients of the two states are close to each other, the considerable overlap of the displacement distributions will lead to difficulties in determining the associated state of a displacement, and consequently large errors in identifying corresponding transition probabilities. SPICER should be especially useful in improving data analysis under this circumstance as it can effectively eliminate confinement error without reducing available data significantly.

To compare the performance of SPICER with traditional $1d$ and $3d$ analyses under these scenarios, we simulated eight different systems with 50,000 trajectories each, with P_{12} and P_{21} set to .0224 ($k = 5/second$) , D_1 to 1 $\mu m^2/s$ and D_2 varied between .8 and .2 $\mu m^2/s$. We analyzed these systems as described and plotted the average percentage errors in D and P of the three methods in Figure 2.7. We also applied the same methodology to the $2d$ data, whose results are shown in Figure 2.8.

Consistent with what we expected, when ΔD decreases, percentage errors in D and P in all three methods increase, but SPICER consistently outperforms the $1d$ and $3d$ analyses, in particular with smaller ΔDs. Only at larger $\Delta D > 0.6$ $\mu m^2/s$ the improvement is less dramatic. These results thus demonstrate SPICER's unique advantage in systems where the diffusion coefficients of two states are closely spaced to each other.

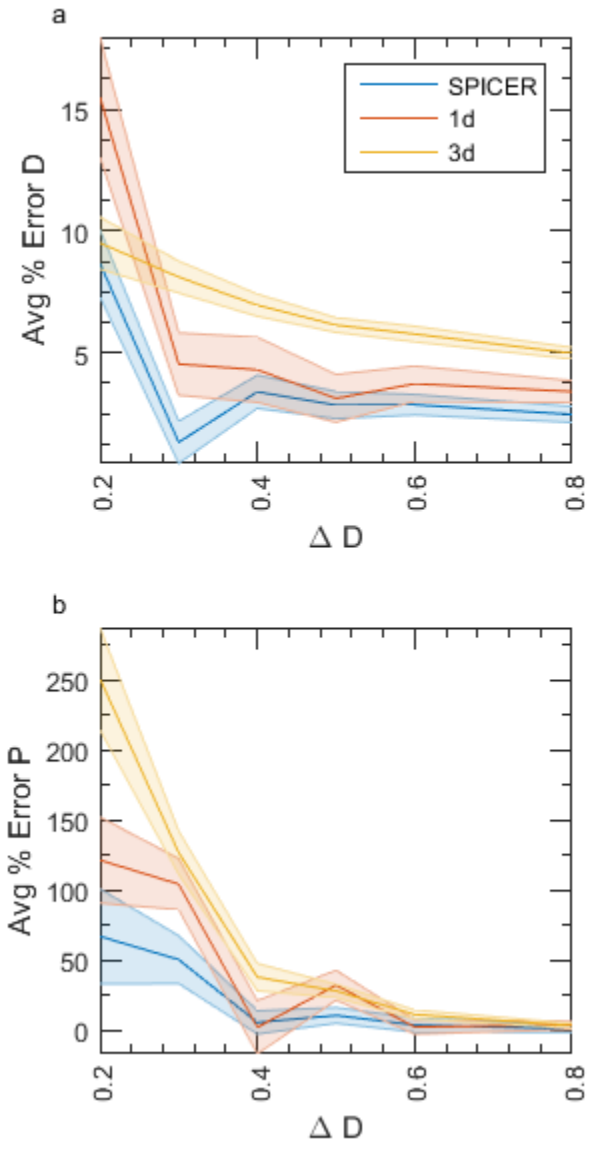

Figure 2.7: Comparison of averaged percent error in identifying diffusion coefficients (a) and transition probabilities (b) of systems with varying separations (ΔD) between the diffusion coefficients of the two states using SPICER, $1d$ or $3d$ analysis. The larger D is fixed at 1 $\mu m^2/s$ with the smaller D varying between 0.8 and 0.2 $\mu m^2/s$. The average percent error is calculated as $(\frac{|D_1-D_1^{true}|}{D_1^{true}} + \frac{|D_2-D_2^{true}|}{D_2^{true}}) \times 50$ or $(\frac{|P_{12}-P_{12}^{true}|}{P_{12}^{true}} + \frac{|P_{21}-P_{21}^{true}|}{P_{21}^{true}}) \times 50$. The shaded region indicates the uncertainty in the parameter and is defined as the standard deviation of the parameter during the MCMC approach.

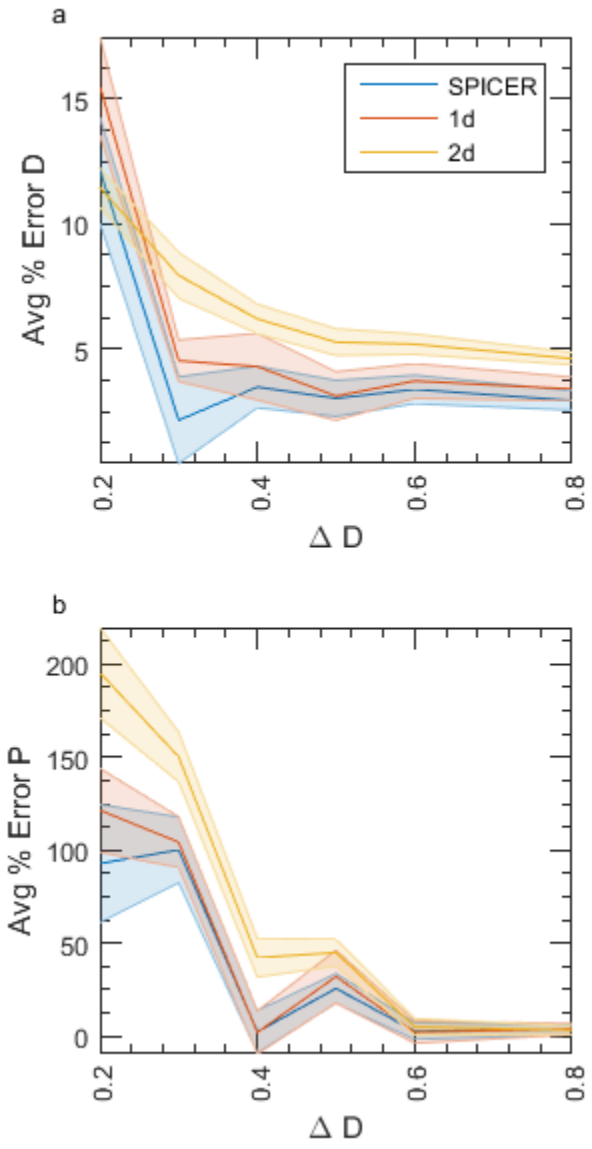

Figure 2.8: Comparison of averaged percent error in identifying diffusion coefficients (a) and transition probabilities (b) of systems with varying separations between the diffusion coefficients of the two states (ΔD) using SPICER, $1d$ or $2d$ analysis. The larger D is fixed at $1\mu m^2/s$ with the smaller D varying between 0.8 and 0.2 $\mu m^2/s$. The average percent error is calculated as $(\frac{|D_1-D_1^{true}|}{D_1^{true}} + \frac{|D_2-D_2^{true}|}{D_2^{true}}) \times 50$ or $(\frac{|P_{12}-P_{12}^{true}|}{P_{12}^{true}} + \frac{|P_{21}-P_{21}^{true}|}{P_{21}^{true}}) \times 50$. The shaded region indicates the uncertainty in the parameter and defined as the standard deviation of the parameter during the MCMC approach.

2.5 SPICER requires a lower number of trajectories to achieve the same level of error reduction compared to $1d$ or $3d$ analysis

SMT analysis usually requires a large number of trajectories (on the order of 10^4 if the average length of trajectories is short [10]), so that diffusion coefficients and state transitions can be determined with statistic significance. However, experimentally it is time-consuming to collect tens of thousands of SMT trajectories. To investigate whether SPICER helps in lowering this requirement, we used the same two-state system analyzed in Figure 2.3 and varied the number of trajectories used in the analysis. In Figure 2.9 we show that for all three methods ($1d$, $3d$, and SPICER) the averaged percent error in D plateaus when the number of trajectories is greater than 5,000; the averaged percent error in P plateaus when the number is greater than 10,000, as accurate determination of P requires a higher number of trajectories. However, at even a low number of trajectories ($\sim$3000), average percent errors of D and P in SPICER are substantially lower than those in $1d$ and $3d$ analyses, approaching the level that would be achieved by 10,000 trajectories with the $1d$ analyses. Note here that the decreases in the total error are mainly brought by the minimization of confinement error in SPICER, which compensates for errors caused by an insufficient number of trajectories, as that in $1d$ and $3d$ analyses. These results demonstrate that increasing the number of trajectories used will not improve the error in the calculated parameters when confinement

error is present in the commonly used $1d$ and $3d$ analyses. The application of SPICER raises the proportion of data without confinement and leads to the least amount of error in determining the diffusion coefficients and transition probabilities in these systems.

2.6 Validating SPICER using experimental RNAP tracking data

To further validate SPICER on experimentally obtained data, we performed $2d$ SMT on RNA polymerase (RNAP) in live *E. coli* cells. RNAP is primarily found within the nucleoid; because of its frequent interactions with chromosomal DNA, it has relatively small diffusion coefficients [40]. Thus, RNAP would experience less confinement from the membrane than other freely diffusing protein molecules in cells, and can serve as a control system with negligible confinement to validate the SPICER algorithm.

We used a functional RNAP-PAmCherry fusion (gift from Dr. Ding J. Jin, National Cancer Institute) that is integrated into the *E. coli* chromosome replacing the endogenous *rpoC* gene, which encodes the β' subunit of RNAP. Under our imaging conditions, we collected a total of ∼25,000 trajectories with the average trajectory length at ∼3 in RNAP-PAmCherry expressing cells grown in minimal M9 medium with a 5-ms exposure time. We first used conventional $1d$ and $2d$ analyses to identify that under this condition the best model describing the diffusive behaviors of RNAP is a two-state model. The

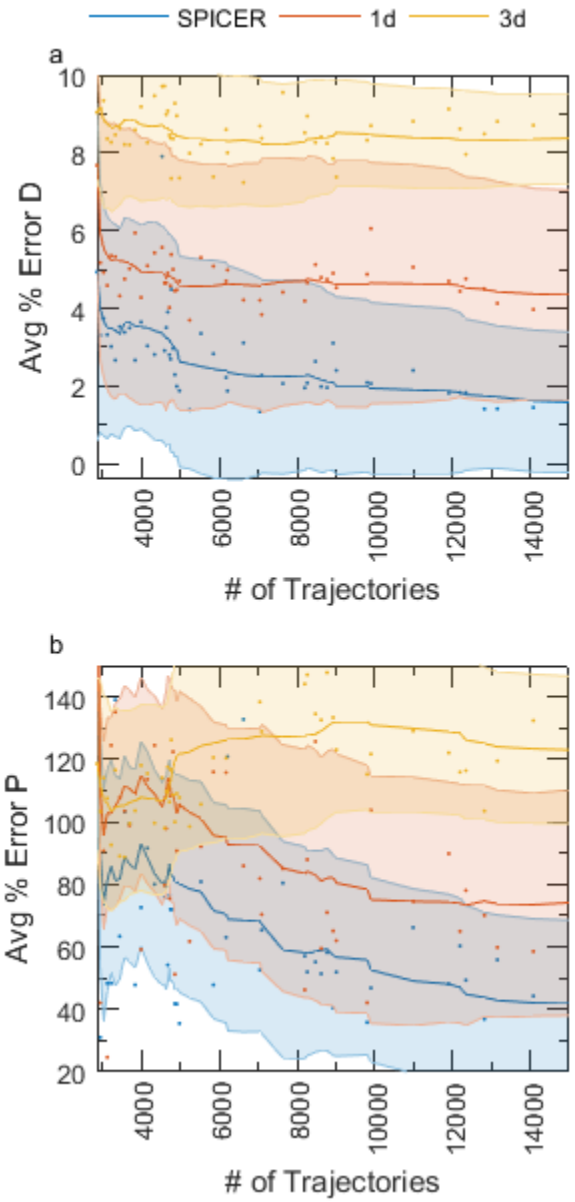

Figure 2.9: Comparison of averaged percent error in identifying diffusion coefficients (a) and transition probabilities (b) of the two state system shown in Figure 2.3C and D with varying number of trajectories using SPICER, $1d$ and $3d$ analyses. Averaged percent error and shaded region are calculated the same way as that in Figure 3. The solid lines are the 10-point moving averages of raw data (scattered dots), and the shaded areas are the moving averages of the standard deviations of the parameters during the MCMC approach.

two D values from $1d$ and $2d$ analyses are similar to each other (D_1 = .38 $\mu m^2/s$, D_2 =.1 $\mu m^2/s$, Figure 2.10A), and are consistent with previous SMT studies of RNAP [40]. However, transition probabilities obtained from the $1d$ analysis are significantly lower than those obtained from the $2d$ analysis (Figure 2.10B). The lower transition probabilities of the $1d$ analysis are most likely due to short trajectory lengths ($\sim$ 3 displacements) combined with reduced amount of data in the $1d$ analysis, which makes it difficult to observe rare transitions between states. Using simulations we further verified that indeed at this slow-diffusion condition, $2d$ analysis describes the system more accurately than the $1d$ analysis (Figure 2.10, C and D).

Next, we applied SPICER using an R-value of 200 nm, identified using the procedure described in the previous section and obtained a new set of D_1, D_2, P_{12} and P_{21}. As shown in Figure 2.10A, diffusion coefficients obtained from the three methods are similar to each other, suggesting that at this slow diffusing rate the confinement error is low and that all the methods are capable of identifying D sufficiently well with the acquired number of trajectories. However, transition probabilities from SPICER and the $2d$ analysis are similar to each other and are both significantly higher than those obtained from the $1d$ analysis (Figure 2.10 B). These results further demonstrate that SPICER can be used to analyze experimental SMT data with high accuracy.

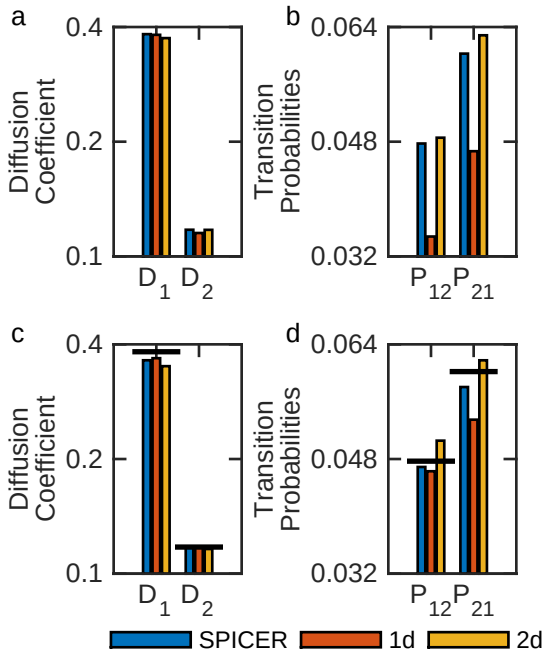

Figure 2.10: Validation of SPICER using experimentally acquired $2d$ SMT data of RNAP in live *E. coli* cells. (a and b): comparison of the identified D_1, D_2 (a) and P_{12} and P_{21} (b) values using SPICER, $1d$ and $2d$ analyses. (c and d) Simulation of a similar system shows the same trend that $2d$ and SPICER analyses are significantly more accurate than the $1d$ analysis, with SPICER reflecting the true values most closely. The true values for the simulation are shown as horizontal black lines.

2.7 Conclusions

In this work we present a simple algorithm, SPICER, to reduce the confinement error in SMT analysis in small bacterial cells. SPICER calculates displacements in all dimensions available ($2d$ or $3d$) and only selectively switch to $1d$ (along the cell's long axis) when a molecule is within a pre-defined R-region where it likely experiences confinement. We provided lookup tables and experimental guidelines for how to find an optimal R-value. The complete package of SPICER is available for download at $github.com/XiaoLabJHU/SPICER$. Using simulations we compared SPICER with commonly used SMT analyses and show that SPICER consistently improves the accuracy in determining diffusion coefficients and state-transition probabilities in SMT analyses. Even when compared to the $1d$ analysis, the traditional method used to relieve confinement in multistate systems, the confinement in the poles of the cells allows SPICER to outperform the $1d$ analysis. This improvement is achieved by increasing the overall proportion of data experiencing free diffusion during the maximization of the likelihood. Furthermore, SPICER performs significantly better than previous methods when the separation in diffusion coefficients of two different states is small, and when the acquired number of SMT trajectories is low ($<$ 3,000). We further validated SPICER using experimentally obtained SMT trajectories of RNAP in live *E. coli* cells. SPICER should be particularly useful for comparing SMT results in bacterial cell size mutants, as the influence of confinement in cells of different sizes can be easily accounted

for in SPICER. Furthermore, the central concept of SPICER can be generalized to other cell geometries as long as localizations in the R-region can be used along a particular dimension in which the molecule experiences the least confinement.

2.8 Methods

2.8.1 Simulations of SMT trajectories with two states

All simulated SMT trajectories used in this work were generated by the software provided in vbSPT using a rod-shaped cell like geometry (unless stated specifically, cell radius = 500 nm and cell length = 2.5 μm) and a single molecule localization error of 20nm [10]. The diffusion coefficients defined in this work take into account this localization error with a time step of 5 ms. The length of individual trajectories follows an exponential distribution with a mean value of 6 steps (each step is 5 ms). The effect of confinement is reflected in the simulation through reflective boundaries at the cell membrane. In the two state model, we assume that State 1 and 2 are defined by two diffusion coefficients D_1 and D_2, and the transition probabilities between them are P_{12} and P_{21}, respectively. The reaction scheme for the maximum likelihood analysis is:

$$D_1 \underset{P_{21}}{\overset{P_{12}}{\rightleftharpoons}} D_2. \tag{2.8}$$

Parameters used in each simulated system in this paper are listed in each corresponding figure.

2.8.2 Likelihood method to identify parameters

We first convert each SMT trajectory to a SPICER trajectory using a fixed R-value. The R-value defines the confinement zone, and is the distance from the membrane boundary of the cell to the edge of the midcell region where the molecule diffuses freely and does not experience confinement. We then take all the converted trajectories and scan the parameter space of the diffusion coefficients D_1, D_2 and transition probabilities P_{12}, P_{21} to obtain the best fit parameters for the system by maximizing the likelihood using a Markov Chain Monte Carlo (MCMC) approach with a preset number of search steps [8]. The MCMC approach begins by selecting a random set of parameters D_1, D_2, P_{12} and P_{21}, and calculates the corresponding summed log likelihood value from all trajectories. A detailed description of the calculation can be found in the section Calculating the Likelihood of Multiple State Trajectories. The process is then iterated by systematically adjusting one of the parameters, chosen at random, by a small amount and then comparing the log likelihood at the new parameter value to the previous log likelihood. If the log likelihood is greater at the new value, the algorithm stays at the new position in parameter space. If the log likelihood is less than the old value, the algorithm takes the difference of the log likelihoods, and two outcomes can happen: 1. If the difference is less than the log of a uniform random number it accepts the new position 2. If the difference is more than the log of a uniform random number,

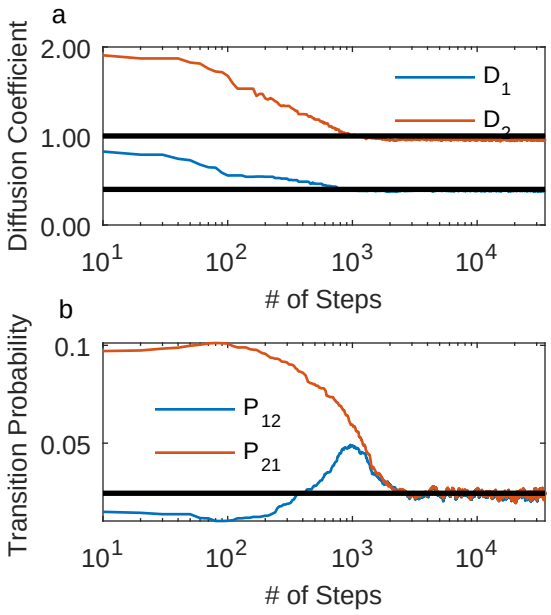

Figure 2.11: An example of a parameter scan using the MCMC approach. The black lines in the two graphs represent the true values, $D_1 = 1\mu m^2/s$, $D_2 = .4\mu m^2/s$, $P_{12} = P_{21} = .0244$ ($k = 5/sec$), of the two state simulation with 50,000 trajectories.

the algorithm stays at the old position. The process repeats by adjusting a new randomly chosen parameter until it reaches a preset number of steps. In all analyses used in this work the number of steps was set at a number large enough so that all the parameters converge well before the end of step numbers.

The stochasticity in the parameter search allows the algorithm to fluctuate around parameters, defining a degree of uncertainty and avoiding local minimums in the parameter search [8]. (An example of a parameter scan on a system is shown in Figure 2.11.) We used the log of the likelihood and summed up the log likelihood of each of the individual trajectories to incorpo-

rate the information from multiple trajectories, see Das *et al.* for the specific algorithms used in this work [8]. The parameters that give the maximum log likelihood are identified as the best-suited parameters for the system. The percent error in this work is defined as $|X_{cal} - X_{true}|/X_{true} \times 100$.

2.8.3 Single molecule tracking data collection and analysis

Single molecule tracking was performed on live MG1655 *E .coli* cells using a photoactivatable fluorescent protein PAmCherry labeled RNA polymerase (RNAP). The PAmCherry gene was C-terminally fused to the *rpoC* gene, which encodes for the β' subunit of RNAP. This fusion gene replaces the endogenous copy in the chromosome, making it the sole source of β' subunit in the cell. Control experiments were performed to ensure that the fusion protein was not subject to proteolytic cleavage, as had been shown previously [137], and that the cells grew otherwise normally as compared to wild-type cells, indicating the functionality of the RNAP fusion.

The RNAP fusion strain was inoculated from a freshly streaked LB plate into 2 mL of minimal M9 media and grown overnight at room temperature, shaking at 250 rpm. After 16 hours of growth, cells were diluted 1:200 into fresh minimal M9 and were shaken at room temperature until they were in mid-log phase growth (OD_{600} of $\sim$0.4). Cells were harvested by taking 1 mL of the cells and spinning them down at 8 rcf for two minutes. Next, 900 μL of

the supernatant was removed from the tube and cells were resuspended in the remaining 100 μL of media, to obtain an OD_{600} of $\sim$4. A small amount of these dense cells, approximately 0.3 to 0.5 μL, was pipetted onto a freshly-prepared 3 % agarose gel pad. Cells were immobilized onto the gel pad by letting the cells dry in air for two minutes. After drying, the gel pad was covered with a clean coverslip to assemble the Bioptechs imaging chamber (Bioptechs Inc.).

Once immobilized on the agarose gel pad, we stochastically activated RNAP-PAmCherry molecules using 0.1 mW of 405 nm light, which converts the PAmCherry molecule from a dark state to a red-emitting state, used 50 mW of 568 nm light to excite individual RNAP-PAmCherry molecules and tracked their cellular positions at a frame rate of approximately 150 Hz (5 ms exposure, 6.74 ms per frame). At this imaging speed, we were able to capture RNAP-PAmCherry molecules up to a diffusion coefficient of 3 $\mu m^2/s$ with accuracy in the cellular position of the molecule of approximately 30 nm. Cellular positions and lengths were determined through the software U-Track [138] and screened based on their intensities and position within the field of view. Trajectories were re-cut into multiple subtrajectories consisting of only consecutive frames of molecular localizations (gaps in localizations are due to the inherent blinking properties of all fluorescent proteins).

2.8.4 Calculating the Likelihood of Multiple State Trajectories:

In this section we describe the methodology created by Das *et al.* to calculate the likelihood of a single particle trajectory with multiple states [8]. For a two state system there are four parameters, $\sigma = [D_1, D_2, P_{12}, P_{21}]$. The likelihood of having a particular single particle trajectory, ω=$(\Delta r_1, \Delta r_2, ...r_N)$, is

$$L(\sigma|\omega) \propto P(\omega|\sigma) = \sum_{All(S)} P(\omega|S,\sigma) \times P(S|\sigma) \tag{2.9}$$

where S is the state sequence of the particle throughout the trajectory, and $All(S)$ is the sum over all of the possible state sequences. The term $P(S|\sigma)$ is the probability of having a particular state sequence S given the two transition probabilities, creating a dependence upon the transition probabilities. The term $P(\omega|S,\sigma)$ is only dependent upon the diffusion coefficients with the particular diffusion coefficient defined by the state sequence S. Because the

summation is over all possible state sequences, we utilize the forward-backward algorithm to calculate the likelihood of a trajectory [8]. The forward-backward algorithm determines the likelihood of a trajectory up to the displacement Δr_j, recursively, using the following equation

$$\alpha_j^i = P[\Delta r_1, \Delta r_2..\Delta r_j, s_j = i|\sigma] = [\alpha_{j-1}^1 * P_{1i} + \alpha_{j-1}^2 * P_{2i}] * P(\Delta r_j|s_j = i, \sigma) \tag{2.10}$$

with

$$P(\Delta r_j|s_j = i, \sigma) = \frac{e^{-\frac{\Delta r_j^2}{4D_i\tau}}}{(4\pi D_i\tau)^{d/2}} \tag{2.11}$$

where α_j^i is the forward variable, which gives the probability of observing the trajectory and being in state i, $s_j = i$ at displacement j. The initial forward variable is calculated from the overall probability of being in either state 1 or 2, see Das *et. al* for details. Given that the total length of the trajectory is N, the probability of having the trajectory ω given the four parameters is

$$l(\sigma|\omega) \propto P(\omega|\sigma) = \alpha_N^{i=1} + \alpha_N^{i=2} \tag{2.12}$$

To account for all trajectories, we calculate the log of the likelihood for each of the trajectories and then maximize the sum of the log of the likelihoods with respect to the four parameters using the MCMC approach as described in Das et al.

$$L(\sigma|\omega_k) = log[l(\sigma|\omega_k)] \tag{2.13}$$

$$L(\sigma|\omega_{All(k)}) = \sum_{k=1}^{M} log[l(\sigma|\omega_k)] \tag{2.14}$$

Chapter 3

Improved Localization Precision in 3D-SMLM Using Weighted Maximum Likelihood Estimation

[1]

3.1 Background

Single molecule localization microscopy (SMLM) relies on the temporal isolation of individual fluorescence emitters to determine the spatial localizations of individual molecules with high precision [7, 5, 6]. SMLM has been widely used in biology to address the spatial organizations and structural dimensions of sub-cellular structures at a resolution ∼10-fold better than the diffraction limit of conventional fluorescence light microscopy [139, 140, 141, 142]. Localization precision, the error in determining the spatial coordinates of a single

[1]Bohrer CH*, Yang X*, Lyu Z, Wang SC, Xiao J. Improved single-molecule localization precision in astigmatism-based 3D superresolution imaging using weighted likelihood estimation. BioRxiv. 2018 Jan 1:304816.

emitter, is a critical parameter of SMLM. Together with sample labeling density [143], localization precision determines the upper bound of achievable spatial resolution [144]. Two-dimensional (2D) SMLM methods can reach a lateral localization precision of 10-40 nm along the x and y dimensions in the focal plane by fitting the single emitter's image to a point spread function (PSF) model. The PSF is usually approximated by using a symmetric 2D-Gaussian function in most algorithms [7, 5, 6]. However, in practice, the true PSF of a given imperfect optical system can deviate significantly from the symmetric 2D-Gaussian function [145].

Recent developments in SMLM have allowed the coordinate of an emitter along the third dimension, the z-axis of the optical path, to be determined with a precision in the range of $15-100$ nm. There are two major approaches [146]: interferometry-based methods such as interferometric photoactivated localization microscopy (iPALM, [147]), and PSF-engineering/extension-based methods such as astigmatism (AS) [148], double-helix (DH) [149], and bi/multi-focal plane (BP) microscopy [150]. Among these methods, iPALM uses the interference of the same photon emitted from an emitter to reach the highest localization precision along the z-dimension at approximatly 15 nm. However, the relatively narrow observation depth (<750 nm above the coverslip, [151]) and complex microscopy setup have limited its broad applications in biology. Multi-focal plane or PSF-engineering methods determine the z-position of single emitters by comparing the image of a single emitter with calibrated PSFs at different z-planes, and in general, can reach a z-resolution in the range of 40-80 nm. Although the z-axis resolution is about 2-3 times worse than the

lateral resolution, these PSF-engineering methods are easy and of low cost to implement. In particular, astigmatism-based 3D-SMLM only requires adding a cylindrical lens in the emission pathway, and hence has seen broad applications within the biological imaging community [152, 148]. In an ideal astigmatism-based optical system, the PSF of a freely rotating single fluorescence emitter can be mimicked by a 2D-Gaussian function [148].

$$PSF(x, y, z) = \frac{1}{(\pi\sigma_X(z_0)\sigma_Y(z_0))} exp(-\frac{(x - x_0)^2}{2\sigma_X(z_0)^2} - \frac{(y - y_0)^2}{2\sigma_Y(z_0)^2}) \tag{3.1}$$

Where the x_0, y_0, and z_0 are the true spatial coordinates of the emitter, and $\sigma_X(z_0)$ and $\sigma_Y(z_0)$ present the widths of the Gaussian function along two perpendicular x and y-axes at z_0. Here, for simplicity, the x and y-axes represent the principle axes of the cylindrical lens respectively. In all astigmatism-based 3D-SMLM imaging, a calibration curve describing the correlation between the astigmatism of an emitter's PSF and its z-position is first established by imaging a fluorescent bead smaller than the diffraction limit at predefined z-planes, and subsequently extracting the astigmatic widths of the emitter's PSF using Equation 3.1. The extracted widths are fitted as a function of the corresponding z-positions using a phenomenological model such as the defocusing function [148] or the quadratic function [153]. By comparing experimentally measured widths of a single emitter with the calibration curves, one can obtain the z position of the emitter with a z-axis localization precision of $40 - 80$ nm under the condition of a nearly perfect optical setup (so that the PSF can be approximated by equation 3.1) [152, 148]. However, in practice, it is labori-

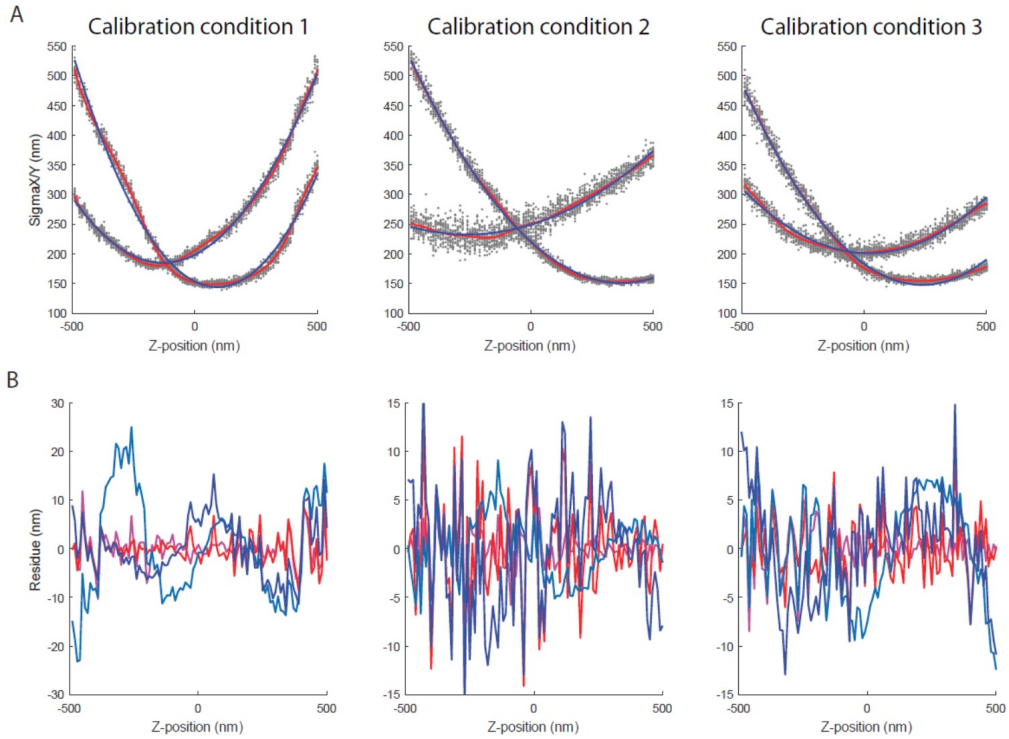

Figure 3.1: A. Quadratic and B-spline fitting of three different ncPSF widths. The optical conditions are corresponding to the three conditions in Fig 3.3. Gray dot: fitted widths from 2D-Gaussian fitting; Red Line: quadratic function fitting of the widths against z; BlueLine: B-spline fitting result. B. Residues of the experimental widths and the fitted curves at different z-planes

ous to perfect the optics each time on a multi-purpose microscope, even for experienced users. As such, imperfect experimental optical setups lead to distorted PSFs and consequently large deviations of measured calibration curves away from these commonly used models, introducing significant uncertainties in determining an emitter's z-position (Fig. 3.1).

To reduce the discrepancy between experimentally measured calibration

curves and the fitting models, Sauer's group used B-spline to interpolate the calibration curves (Fig. 3.1), which was able to obtain higher accuracy and flexibility than the original fitting functions under various experimental conditions [154]. Additionally, Shaevitz et al. used a Bayesian interference method to measure the probability distribution of astigmatic PSF widths at different z-positions [155]. Nevertheless, these methods still assume that each emitter's PSF can be approximated by an elliptical 2D Gaussian function, which doesn't necessarily hold true in an imperfect optical system (Figure 3.2A). For instance, the maximum intensity position of the PSF can shift, or "wobble," due to coverslip-tilt and non-rotational symmetric aberration of an individual objective or other components in microscope [156]. Additionally, spherical and other aberrations can distort the PSF shape, which introduces bias and compromises the localization precision in the z-position as well as in x-y (Figure 3.2A) [145] .

An analytical description of the PSF, under specific conditions, can be retrieved from the pupil function of the imaging setup and has been shown to improve z-axis resolution [157]. The pupil function is then interpolated or decomposed in Zernike polynomials to calculate the PSF at different z-positions. The 3D phase retrieval (PR) method has been implemented in BP and DH microscopy to approximate the pupil function [157]. However, these PDF-retrieval methods are tedious to implement and require a thorough understanding of the optical setup.

In this chapter, we introduce a different approach: using the experimentally measured PSF and a numerically weighted maximum likelihood estima-

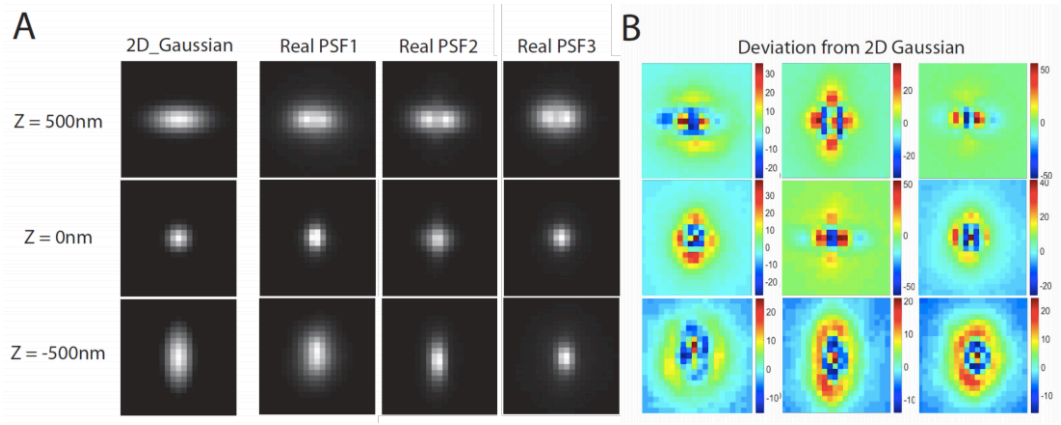

Figure 3.2: Deviation of experimentally measured ncPSF from a 2D-Gaussian model. (A) Simulated 2D-Gaussian PSF image (first column) and TetraSpeckTM beads images (experimental ncPSF) at different z-planes (-500, 0, 500 nm) with three different optical setups. PSF2 is adjusted from PSF1 by shifting the cylindrical lens position along the optical axis while PSF3 is by changing the correction collar position of the objective. (B) Numerical deviation of experimental ncPSFs from the corresponding best 2D-Gaussian fittings.

tion (WLE) to improve the z-axis localization precision of single emitters in astigmatism-based SMLM. Our method is based on the principle that the PSF of an emitter at different z-positions can be characterized as an experimentally measured image independent of any a priori model assumptions such as an elliptical Gaussian model. For each experimentally measured image of an emitter, our method numerically determines the probability for each pixel to have a particular signal level given its z-position using the image of a calibration bead at each z-plane with an experimentally characterized background noise distribution. We then weight the importance of the pixels of an emitter by conducting a phase space search to minimize the calculated z-distances between repeated localizations from the same emitters. We verified that by maximizing the weighted likelihood, we could reach a comparable or higher localization precision when compared to the B-spline based traditional Least Square (LS) fitting methodologies for 2D Gaussian PSFs, independent of the optical setup. It shows the highest improvement when the experimentally measured PSF deviates significantly from the standard elliptical 2D Gaussian model. Thus, the WLE approach alleviates the practical concerns in perfecting the optical alignment and enables improved z-axis localization in astigmatism-based 3D superresolution imaging.

3.2 Principle and workflow of WLE

To estimate an emitter's coordinates, least square fitting (LS) and maximum likelihood estimator (MLE) based on a known PSF, are the most com-

monly used methods. LS is computationally faster than MLE and has produced comparable precision in experiments where high photon counts are achievable. In SMLM fitting algorithms, LS fitting is often chosen over MLE to increase the computational speed and simplify the fitting process. However, when the signal to noise ratio (SNR) is low, LS leads to a compromised localization precision. MLE is more computationally intensive compared to LS, but with a correct PSF function and a correct noise model [158], one can theoretically approach the upper bound of the estimation precision, the Crame-Rao limitation [159].

Our approach, termed weighted maximum likelihood estimation (WLE), determines a single emitter's z-position by maximizing the weighted likelihood of having a particular signal for each pixel at a particular z-position according to an experimentally determined probability density matrix (PDM). WLE is very similar to the MLE approach, but it assigns the information from different sources, different pixels in this case, varying degrees of (weighted) importance. The PDM was determined by convolving the numerically calibrated point spread function (ncPSF) with scaled photon noise and the background noise distribution. In essence, WLE finds a single emitter's z-position by numerically matching its experimentally measured image with that of a calibration bead at a known z-position considering the intensity and background distribution of each pixel. Below, we describe the five main steps to implement the WLE algorithm (Fig. 3.3).

In the first step, we estimated the ncPSF by imaging a bright fluorescent bead (100 nm in diameter) at a series of evenly spaced, 10-nm apart, z-planes

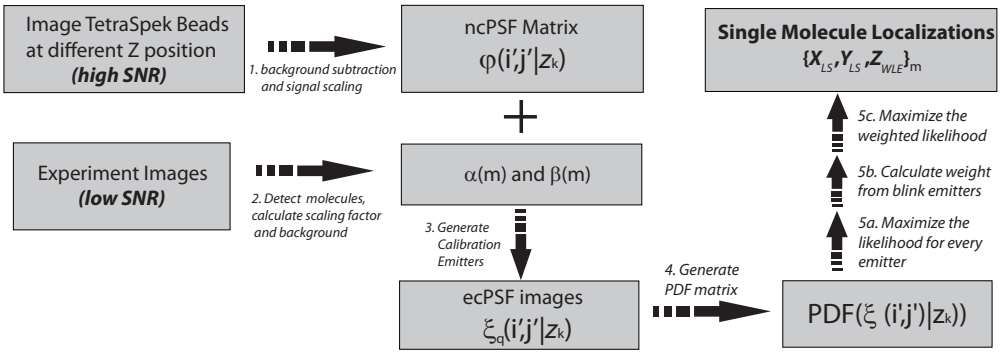

Figure 3.3: Schematics of the WLE workflow. Experimentally measured bead images at different z-planes and real emitter images were used for localization.

using a piezo-stage. As a point source, the background-free averaged and normalized image represents the ncPSF of the optical system. To increase the signal to noise ratio, we took multiple (N) images at each z-plane. We then computed the ncPSF, $\varphi_{ij}(k)$, for each kth z-plane as:

$$\varphi_{ij}(k) = \left\langle \frac{I_{B,ij}(n,k) - \beta(n,k))}{\sum_{i,j}[I_{B,ij}(n,k) - \beta(n,k)]} \right\rangle \tag{3.2}$$

where $I_{B,ij}(n,k)$ is the intensity of the pixel at row i and column j in the nth bead image (21 by 21 pixels in our example and can be varied) at the kth z-plane, and $\beta(n,k)$ is the background intensity calculated using averaged intensity values of pixels furthest away from the bead center. The mean background intensity $\beta(n,k)$ was subtracted from $I_{B,ij}(n,k)$ to obtain the true signal intensity at each pixel, which was further normalized by the integrated signal intensity of the image after background subtraction. The final ncPSF was obtained by averaging over all N images of the bead for the kth z-plane.

This estimation of ncPSF is sufficient because of the negligible noise level due to the high signal from fluorescent beads and averaging over N images allows us to approximate the final background of the mean image as zero (Methods).

In the second step, we obtained images containing single-molecule emitters from simulation or experimentally measured astigmatism-based imaging (Methods). A wavelet-filter based algorithm was applied to identify and crop out the local maxima into regions of individual emitters with the same size as ncPSF. We then calculated the background noise distribution, β, using the peripheral pixels in the cropped images. Here, we assumed that the noise distribution was identical among the pixels in all cropped images, which can be further adjusted depending upon whether the background is uniform among pixels. To obtain the PDM, we then determined the scaling factor $\alpha(m)$ with the following:

$$\alpha(m) = \sum_{ij} [I_{E,ij}(m) - \langle \beta \rangle] \tag{3.3}$$

where $I_{E,ij}(m)$ is the intensity of each pixel for the cropped emitter m. Therefore, $\alpha(m)$ is proportional to the total photon number from each single emitter. Here we assumed that the distribution of the scaling factors among different emitters was independent of their z-positions. We validated this assumption using experimental data (Fig. 3.4).

In the third step, we determined the PDM, which allowed us to calculate the weighted likelihood. We utilized the ncPSF, $\varphi_{ij}(k)$, $k = 1, 2, ...$ estimated in Step 1 to generate q calibration emitters at each z-plane k, $(\xi_{ij}(q, k))$, incorporating the previously determined background noise β and scaling factors

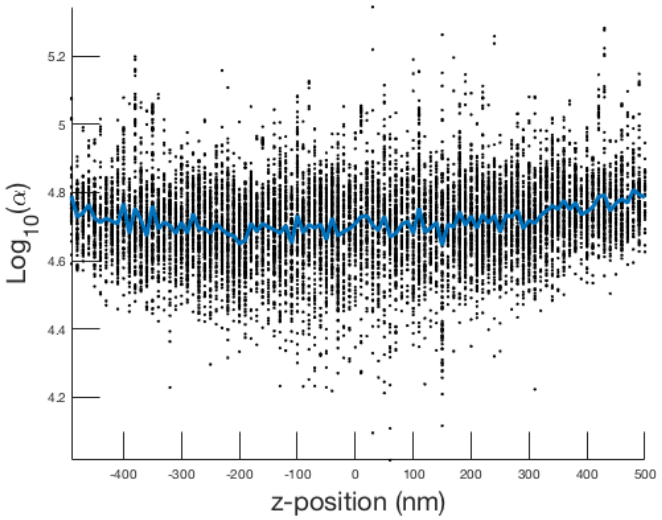

Figure 3.4: The α values of individual emitters at various z-planes, black dots, with the mean value of each all the emitters in each z-pane displayed as a blue line. The α values of each emitter were determined by first subtracting the mean background of each cropped emitter and then summing over all pixels.

α. Note here that β and α are randomly sampled variables from their corresponding distributions (Methods). To generate the calibration emitters, we first simulated a signal for each pixel using the ncPSF multiplied by the scaling factor α with Poisson noise (photon noise) in each pixel: $Poisson[\varphi_{ij}(k) \times \alpha]$. Here we assumed Poisson noise for simplicity, but other circumstances could be accommodated. The background noise of each pixel was added by randomly sampling the noise distribution β. For simplicity and computational ease, we generated 4000 calibration emitters at each z-plane with varying α and background noise. In principle, one could instead generate the calibration emitters at each z-plane for each specific α and incorporate the specific background noise distribution of an emitter's cropped pixels, which will likely increase z-axis localization precision even further. Here the experimentally calibrated emitter image, which we named ecPSF, is:

$$\xi_{ij}(q,k) = \frac{Poisson[\varphi_{ij}(k) \times \alpha] + \beta}{\alpha} \tag{3.4}$$

Here α, the scaling factor, is randomly sampled from its distribution and has the same value in both the denominator and numerator of the equation. We then linearly shifted the centroids of the ecPSFs (estimated using a 2D-Gaussian fitting) so that the "centers" of all the adjusted PSF's were aligned at each z-plane.

Next, we approximated the probability density distribution to observe the signal, Λ at pixel (i',j') for the k^{th} z-plane with $\xi_{i'j'}(q,k)$: $\Psi(\Lambda|i',j',k) \approx PDF(\xi_{i'j'}(1..Q,k))$, where PDF is the approximated probability density func-

tion for the term within the parentheses and Q is the total number of experimental calibration emitters for that z-plane (Methods). Here $\Psi(\Lambda|i', j', k)$ is what we referred to as the PDM, the probability density matrix, which we then used below to calculate the likelihood.

In the last step, for each single emitter image, we normalized the intensity to the total intensity of each image after background subtraction and determined the centroid as in the previous steps. This procedure resulted in the adjusted signal $\Lambda_m(i', j')$ for the m^{th} emitter. We then determined the optimal z-position for the m^{th} emitter by maximizing the following: $L(k) = \Sigma\omega(i', j') \times log(\Psi(\Lambda_m(i', j')|i', j', k))$ for the k^{th} z-plane. Here the elements of ω contain the weighted importance for each pixel (Methods, Fig. 3.5). We determined the weights that resulted in the best resolution by performing a phase space search, adjusting each element of ω to minimize the distance between repeated localizations of the same emitters (Methods, Fig. 3.6). (We provide a user guide, code and example data allowing one to implement and understand the inner workings of WLE ($https://github.com/XiaoLabJHU/WLE$).)

3.3 Validation of WLE

To validate the WLE algorithm, we imaged TetraSpeck^{TM} fluorescence beads on a coverslip scanning 100 z-planes at 10-nm intervals. For each z-plane, we acquired 200 images and observed minimal photobleaching. For astigmatism-based 3D SMLM imaging, it is necessary to adjust the objective correction collar and the position of the cylindrical lens to minimize the spherical aberration caused by refractive index mismatch and the cylindrical lens

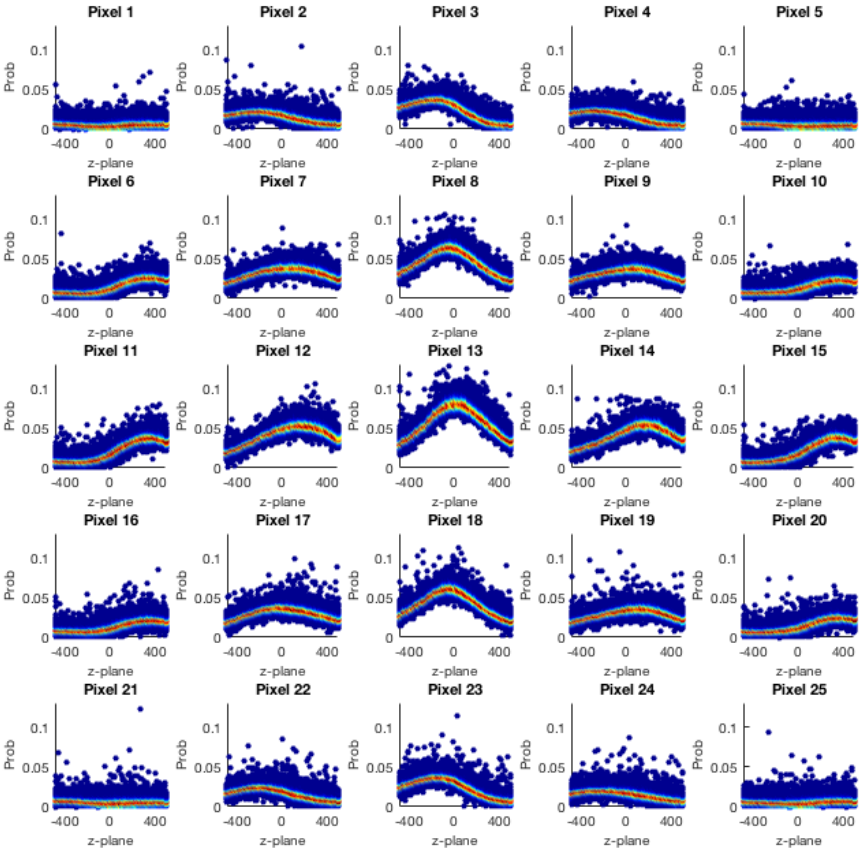

Figure 3.5: The probability to have a particular signal in each pixel for each z-plane, for a high signal to noise ratio for experimental ncPSF 2. The distribution of Pixel 13 provides a much greater amount of information than the distribution in Pixel 25.

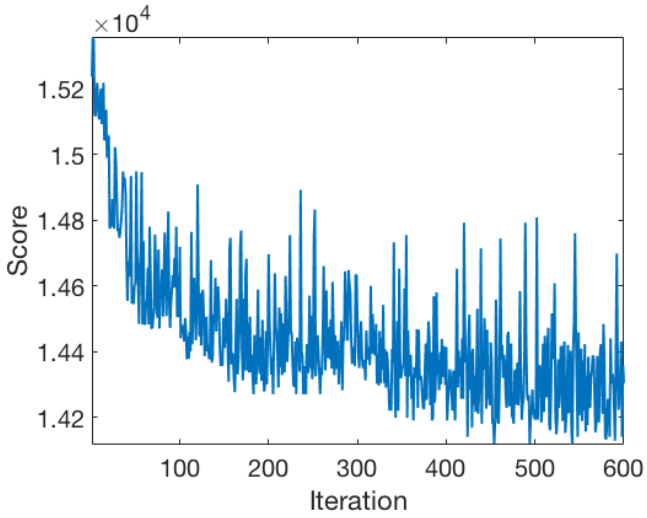

Figure 3.6: Converging of the Score function for the experimental ncPSF 2 versus iterations. We stopped the phase space search after the Scoring function reached a plateau. On average the convergence of the Score function led to an improvement in the resolution by 5nm and varied depending upon the condition being analyzed.

itself. We mimicked these adjustments and obtained three sets of calibration PSF images at different settings (PSF I, II and III, Fig. 3.2 A). All these experimentally measured PSFs deviated from the perfect 2D-Gaussian function (Fig. 3.2 B). Nevertheless, as a comparison, we fit these PSFs using Equation 3.1 to obtain the centroid positions (x_0, y_0) and the astigmatic PSF widths, $\sigma_X(z_0)$ and $\sigma_Y(z_0)$, at various z-planes (Fig. 3.1 A). As shown in Fig. 3.1 A, although the B-spline function fit the correlation between the widths and z-plane significantly better than a quadratic function, the correlation shape and the corresponding errors varied dramatically for the three different conditions, indicating high levels of uncertainty introduced by LS-based 2D-Gaussian fitting due to distorted PSF shapes.

To evaluate the performance of WLE in comparison with LS-based B-spline and quadratic fitting methods, we simulated single emitters for each of these experimental PSFs at various z-planes with different signal to noise ratios (SNRs) (Fig. 3.7 , Table 3.1 and Methods). For LS-based B-spline and quadratic fitting methods, we fitted images of single emitters from highest SNR data (Fig. 3.10) with Equation 3.1 to obtain the z-positions using different calibration functions (B-spline or quadratic (Fig. 3.7). Here we defined the error, or the localization precision, as the mean absolute distance of all emitters from their true locations under each condition. As shown in Fig. 3.7 , for all three different optical settings and at different SNR, the quadratic function (purple) performed most poorly and reached a plateau of error at $\approx$ 40 to 60 nm for PSF I and III. The B-spline method was significantly better than the quadratic method and was able to reach a maximum resolution of $\approx$10 nm

with the highest SNR for all three PSFs, suggesting that it is more reliable and adaptive in fitting the calibration curve compared to the quadratic function, as shown previously [154].

For WLE, we determined the z-position of each simulated emitter by comparing its corresponding image with the bead-generated ncPSF and maximizing the weighted likelihood. For PSFs I and II, we found that WLE resulted in similar localization precisions compared to B-spline when SNRs were low but showed more significant improvement than B-spline when the SNR was high. For PSF III, we observed the greatest improvement (≈1.5-fold) by WLE compared to B-spline, reaching a localization precision of < 10 nm and surpassing all other methodologies for every SNR we tested. These results illustrated that WLE consistently performed equally well or better than the best-performing LS-based B-spline method under all tested conditions.

3.4 WLE improved z-axis localization precision independent of PSF shape

Next, we reasoned that the varied levels of improvement of WLE over B-spline (Fig. 3.7 A to C) could be due to the deviation of the experimentally measured PSFs from an ideal 2D Gaussian function – the larger the deviation, the better WLE outperforms the LS methods. Therefore, we quantified the deviation of a PSF from an ideal elliptical 2D Gaussian as the mean of the absolute difference between the bead images and a 2D Gaussian fit to the bead images at each z-plane. We then plotted the average localization precisions at different z-planes by the three methods against the PSF deviation values. As

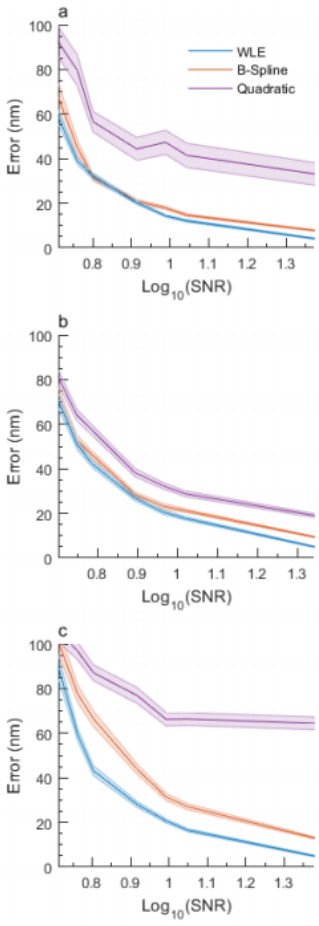

Figure 3.7: Average localization precision of LS-fitting (purple for quadratic and orange for B-spline) and WLE (blue) using synthetic images generated from experimental ncPSFs (Figure 3.4) at a series of signal to noise ratio (SNR).

shown in Fig. 3.8 , we observed significant correlations between the localization precision and the Gaussian deviation value for the quadratic and B-spline based LS fitting methods – the larger the deviation, the worse the localization precision. In contrast, WLE showed no correlation, and the determined localization precision stayed essentially flat across the different Gaussian Deviation values (Fig. 3.8 , blue). The quadratic method sometimes even showed low precisions at low deviation PSF conditions (PSF I), which resulted from the significant deviation of the $\sigma_X(z_0), \sigma_Y(z_0)$ from the quadratic calibration function for PSF I. In particular, we observed the largest Gaussian Deviation values (>110) for PSF III, for which the corresponding localization precisions determined by the B-spline and quadratic methods degraded dramatically, whereas the WLE error remained approximately constant. In contrast, when we applied WLE and B-spline to an ideal elliptical 2D-Gaussian PSF, we obtained equally good localization precisions for both methods across different SNR conditions (Fig. S5, Methods). These results strongly suggested that the distortion of PSF shape from the ideal 2D-Gaussian caused by imperfect optical setups was a major factor leading to the low localization precisions in LS fitting-based localization, and that the WLE-based localization is independent of the shape of PSF.

3.5 Discussion

In this work, we developed a WLE methodology to enhance the localization precision along the z-axis in astigmatism-based SMLM without an analytical description of the PSF. We validated WLE by analyzing simulated data using

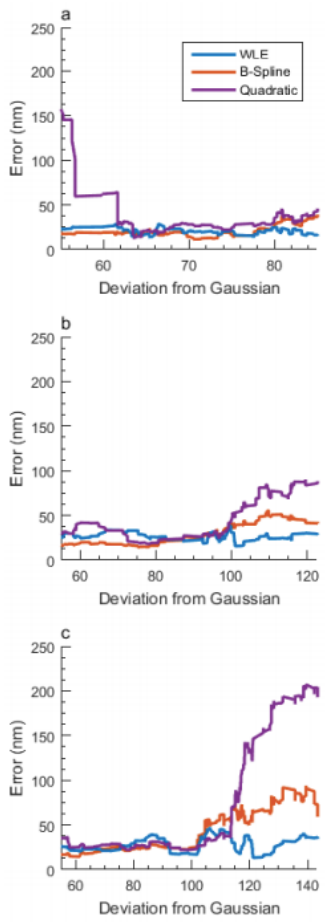

Figure 3.8: The average localization precision of LS-fitting (purple for quadratic and orange for B-spline) and WLE (blue) at different deviations of the PSF from a 2D-Gassian model.

three experimentally measured PSFs. We found that WLE resulted in similar localization precision when compared to the commonly used B-spline fitting method when the PSF was approximately Gaussian. WLE surpassed B-spline significantly when the PSF deviated from the ideal shape, which is likely the case for real-world astigmatism-based SMLM experiments.

The major advantage of the WLE method is that it does not require a specific predefined PSF model or a user-defined noise distribution model. Both can be experimentally measured and utilized to generate calibration images, which allows WLE to predict an emitter's z-position by numerically "matching" its image with weighted pixels to calibration images. WLE does not require the calculation of a complex pupil function, which facilitates its use by non optics-oriented users. Because of its independence of PSF shape, WLE can also be applied to BP or other PSF engineering based 3D SMLM methods. An additional novel aspect of WLE is to determine the importance of the information in each pixel (the weight) in an unbiased manner by minimizing the z-distance difference between repeated localizations of the same emitters.

The current WLE algorithm does not correct the PSF change caused by the refractive index mismatch. Using an experimentally measured PSF in different z-planes away from the cover glass such as fluorescence beads on another inclined surface would solve this problem [160]. The major disadvantage of WLE is that it is computationally intensive because each emitter's image needs to be analyzed independently. Currently, the construction of a 3D superresolution image of $\sim$2000 molecules takes $\sim$150 CPU hours, but the current code can be further optimized for speed. With parallel or GPU computation we foresee

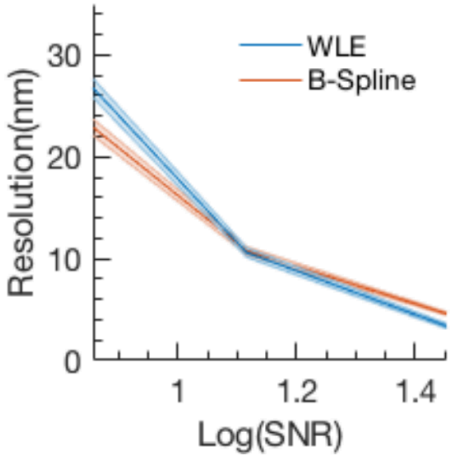

Figure 3.9: The results of applying the WLE and B-Spline methodologies to a perfect 2D Gaussian PSF. Where the shaded areas represent 1 SEM determined by bootstrapping.

significant improvement of the computational speed of WLE.

3.6 Methods

3.6.1 Experimental Methods

We imaged TetraSpeck^{TM} fluorescence beads (ThermoFisher Scientific, T7279) sparsely distributed on a coverslip in Phosphate-buffered Saline (PBS) with a 1.49 NA oil immersion objective (100X, Olympus). Using an ASI piezo stage we collected 200 images per z-plane, for 100 z-planes in 10 nm intervals (1 μm range). We then cropped out a 21 by 21 pixels region containing only a single bead to generate the calibration images at different z-planes. To obtain different PSFs at different optical settings, we adjusted the correction collar of the objective and the position of the cylindrical lens along the emission path.

These slight modifications happen routinely on a multi-user microscope.

3.6.2 Simulating Experimental PSFs and Images

Once we had acquired the images of the beads in each optical setup, we used a cubic-spline function to interpolate the images to obtain pseudo-subpixel images with 10nm in pixel size (original pixel size was 160 nm). We then removed the mean background intensity counts and further normalized these images by the total intensity to generate the ncPSFs at different z-planes. Based on these ncPSFs, we simulated a series of single emitter images with different signal to noise ratios (SNR) at various z-planes, with each emitter's centroid at the center of the images. For each simulation condition, we pre-determined the total signal level (I_s) and background level (I_{bn} and I_{bp}) in number of photons (Fig. 3.10). The subpixel PSF was then multiplied by the total photon number of a single image that was generated from a Poisson distribution with the expectation I_s. We then added Poisson noise to each small pixel to mimic the photon noise. The final signal image was generated by summing the photons in neighbor pixels to recover the experimental pixel size (160 nm). The background was generated directly in an image with the experimental pixel size. It contained three parts: Gaussian noise (I_{bn}) to mimic the instrumental noise or readout noise, Poisson noise (I_{bp}) to mimic light-scattering noise plus dark current, and a constant background offset. Given the independence of the signal and background in an ideal experiment, we added these two images to produce the simulated experimental images.

Table S1. Simulation Parameters and Signal to Noise Ratio

Simulation Condition	I_s [a]	I_{bn} [b]	I_{bp} [c]	offset	SNR[d] (PSF)		
					1	2	3
1	250	1	3	10	14.96	15.03	14.45
2	375	1	3	10	18.24	18.42	17.45
3	500	1	3	10	21.51	21.78	20.27
4	1000	1	3	10	34.22	34.91	31.54
5	1500	1	3	10	46.44	47.39	42.21
6	2000	1	3	10	57.88	59.14	52.55
7	10000	1	0.6	10	228.89	232.97	208.53

a. The average total signal per image used in simulation.

b. The average normal distributed noise per pixel.

c. The average Poisson distributed noise per pixel.

d. The SNR is calculated by averaging single image SNRs over all z-planes. The single image SNR is defined as the maximum pixel intensity in each image subtracted the mean background and then divided by the average noise (standard deviation of the peripheral pixels).

Note: in experimental PALM image (fluorescent protein), the SNR is in the range of condition 1-3 while the STORM image (organic dye) usually has higher SNR (condition 4-6).

Figure 3.10: Simulation Parameters and Signal to Noise Ratio

We repeated this process to obtain 200 images at 100 different z-plane as our calibration images, but with different SNRs. We calculated SNR by dividing the highest pixel intensity by the average noise (sum of photon noise and background noise) for each pixel.

3.6.3 Logic for the Scaling of the Bead PSFs and Incorporating Poisson Noise

In this work we assumed that the image we obtained from a single emitter with the z-position, Z_k, is composed of two individual entities, the signal distributed with the underlying deterministic function, ncPSF, and a background term

randomly sampled from a noise distribution for each pixel i and j of each cropped image:

$$Image(i,j|Z) = \eta \times P(i,j|\eta, PSF(i,j|Z)) \times PSF(i,j|Z) + \beta, \tag{3.5}$$

where η scales the first term that makes up the observed signal from the emitter and β is a random variable that does not vary between pixels and is independent of the characteristics of the emitter (Uniform Background). The photon noise, who's distribution is dictated by $P(i,j|\eta, PSF(i,j|Z))$, is likely dependent upon both η and $PSF(i,j|Z)$. For instance, the first term, $\eta \times P(i,j|\eta, PSF(i,j|Z)) \times PSF(i,j|Z)$, can be approximated as a Poisson distribution, $Poisson[\eta \times PSF(i,j|Z_k)]$. Thus the characteristics of the Poisson are dictated by both terms η and $PSF(i,j|Z)$.

To obtain the probability density matrix (PDM) for the observance of an emitter in the real experimental sample we first obtained an approximation for the deterministic component of the signal, $PSF(i,j|Z)$, from the bead sample by taking an average over all the images from the same z-plane:

$$\langle \frac{Image_{bead}(i,j|Z) - \langle\beta\rangle}{\eta \times P(i,j|\eta, PSF(i,j|Z))} \rangle =$$

$$\frac{\langle Image_{bead}(i,j|Z) - \langle\beta\rangle\rangle}{\langle \eta \times P(i,j|\eta, PSF(i,j|Z))\rangle} =$$

$$\frac{\langle \eta \times P(i,j|\eta, PSF(i,j|Z)) \times PSF(i,j|Z) + \beta(i,j) - \langle\beta\rangle\rangle}{\langle \eta \times P(i,j|\eta, PSF(i,j|Z))\rangle} =$$

$$\frac{\langle \eta \times P(i,j|\eta, PSF(i,j|Z))\rangle \times PSF(i,j|Z) + \langle\beta(i,j)\rangle - \langle\beta\rangle}{\langle \eta \times P(i,j|\eta, PSF(i,j|Z))\rangle} =$$

$$\frac{\langle \eta \times P(i,j|\eta, PSF(i,j|Z))\rangle \times PSF(i,j|Z)}{\langle \eta \times P(i,j|\eta, PSF(i,j|Z))\rangle} + \frac{\langle \beta(i,j)\rangle - \langle \beta\rangle}{\langle \eta \times P(i,j|\eta, PSF(i,j|Z))\rangle} =$$

$$PSF(i,j|Z) + \frac{\langle \beta(i,j)\rangle - \langle \beta\rangle}{\langle \eta \times P(i,j|\eta, PSF(i,j|Z))\rangle} \approx PSF(i,j|Z) \approx$$

$$\langle \frac{Image_{bead}(i,j|Z) - \langle \beta\rangle}{\sum_{ij}[Image_{bead}(i,j|Z) - \langle \beta\rangle]} \rangle$$

Where $\langle \beta \rangle$ is the mean of the background, which is assumed to be small compared to the signal in the TetraSpek^{TM} bead sample as η is large, and $\langle \beta \rangle$ considered to be well defined and approximately uniform, causing the remains of the background terms to approach zero, $\frac{\langle \beta(i,j)\rangle - \langle \beta\rangle}{\langle \eta \times P(i,j|\eta, PSF(i,j|Z))\rangle} \approx 0$.

Second, in real experiments, the integrated signal (also the scaling factor $A_m,\ m = 1, 2, .., M$) of each emitter is not a constant. For simplicity, we constructed an experimental distribution from the signal distribution of all the M emitters:

$$PDF_{scaling}(\alpha) = \lim_{\Delta\alpha \to 0} (\frac{1}{\Delta\alpha} \frac{\{\alpha < A_m < \alpha + \Delta\alpha\}}{\{A_m\}}) \tag{3.6}$$

Here, α is the scaling factor of continuous random variable while A_m is the experimental measured scaling factors from emitter m. We calculated the signal of every detected emitter without background by subtracting the mean background for the emitter, which was estimated by taking the average of the mean intensity of the peripheral pixels.

On the other hand, we assumed the noise is uniformly distributed and independent on the signal. The probability density of the background, hence,

can be sampled from the experimental images as:

$$PDF_{backgound}(\beta) = \lim_{\Delta\beta \to 0}(\frac{1}{\Delta\beta}\frac{\{\beta < B_l < \beta + \Delta\beta\}}{\{B_l\}}), \tag{3.7}$$

where β is the random variable and B_l is the measured single pixel background as described in the scaling factor part. With a certain scaling factor, PSF, and noise distribution, we could generate the PDM of the single pixel intensity as:

$$PDM(Image_{em}(i,j|PSF(Z_k), A(1...m), B(1...l))) \approx$$

$$PDF(\frac{Poisson[\alpha \times PSF(i,j|Z_k)] + \beta}{\alpha}).$$

The conditional probability of the intensity at pixel i and j of a certain emitter is related to the PSF shape at Z position convolved with the signal distribution and the noise distribution. The two approximations we used here are: (1) the photon noise in each pixel is Poisson; (2) the background noise is additive to the signal. Practically, we generated this matrix by simulating the experimental calibrated emitters as described in the main text. This equation allowed us to calculate the likelihood to have a particular signal by approximating the distribution to have the altered signal by simulating the signal at the various z-planes. More specifically, we approximated the PDM of the signal by setting the number of bins so that the probability histogram of each pixel at each z-plane of the altered signal had a maximum value of 15%. We then normalized it by the bin widths to create the PDM.

3.6.4 Justification that the Background Scaling Factor Is Independent of Z-plane

To determine whether the distribution of α (the total photon numbers from the single emitters within the actual experiment) was independent of z-position, we imaged immobilized AlexaFluor 647 dyes on coverslips at different z-planes. The molecules could blink multiple times and be recorded. Therefore, we can measure the A_m at various z-planes. The results of the experiment are shown in Fig. 3.4 with the mean A which is the expectation of α of all emitters at each z-plane in blue and A_m values for each emitter shown in black (Please note the Log scale). These results demonstrate that there is a similar amount of variation within the α values among emitters across all z-planes and the mean shows very little variation at the various z-planes, justifying the assumption that α is independent of z-plane (Figure 3.4).

3.6.5 Brief Description of Weighted MLE and Motivation

(Following the main text description in the section Principle and workflow of WLE) To determine the z-plane of each individual emitter we maximized the following:$L(k) = \sum \omega(i', j') \times log(\Psi(\Lambda_m(i', j')|i', j', k))$. Here the elements of ω contained the weighted importance for each pixel. Now the question is why do we add in the weights and not just utilize the traditional MLE. This can be thought of as each pixel is a different type of observation. Since the intensity decreases fast outward of the central position, incorporating more pixels far away from the center could potentially cause the noise to dominate the maximization process, as more information does not always convey more

predictive strength (Assuming error in defining the probability distributions). To illustrate the variation of the signal with z-plane, Figure. 3.5 below shows the distributions of having a normalized signal at a z-plane for the pixels of aligned cropped emitters. The methodology used to determine the importance of each pixels is discussed in the following section.

3.6.6 Methodology to Determine the Weights $\omega(i', j')$

To determine the importance of each pixel we sought to minimize the following scoring function for emitters that are within 200 nm of each emitter (determined from their x and y coordinates),

$$Score = \sum_{i}^{n} \sum_{k}^{n} f(i,k),$$

$$f(i,k) = |Z_i - Z_k|, if |Frame(i) - Frame(k)| < 5$$

$$f(i,k) = 0, if |Frame(i) - Frame(k)| > 5.$$

Where Z_i and Z_k are the calculated z-positions of the emitters i and k with the total number of emitters equal to n and $Frame(i)$ and $Frame(k)$ are the frames of the two emitters. We then used the following simple algorithm to conduct a phase space search to minimize Score. An example of the Score function for the lowest SNR experimental PSF II is shown in Figure. 3.6.

First step: we start with all of the elements of $\omega(i, j)$ equal to 1, this is equivalent to the more traditional MLE approach.

Second step: we then calculate the z-positions of all emitters by maximizing the weighted likelihood function, as discussed in the main text.

Third step: we then calculate Score and if it reaches a minimum (of the past Score values) we store $\omega(i, j)$ and the calculated z-positions of the emitters.

Fourth step: we then randomly choose and element of the stored $\omega(i, j)$ and replace it with a new random uniform random number.

Fifth step: we then go back to the second step and proceed until the number of iterations is reached. (The Score function should have converged on a minimum aka. reached a plateau) In Figure. 3.6 we show how the Score function converges for one of the simulation datasets.

3.6.7 Application of Fitting Methodologies to Perfect Gaussian PSF

Here we sought to determine how the WLE methodology compared with the B-spline methodology if the PSF was a true Gaussian with a background noise model that optimized the performance of B-spline. To do this we simulated a perfect Gaussian PSF with the background noise described in the previous sections at three different SNRs. We then applied WLE and the B-spline methodologies, the results are shown in Fig. 3.9 Here the B-spline methodology surpassed the WLE methodology at the lowest SNR and was approximately equal to WLE at the other SNRs. We believe the B-spline methodology was able to surpass WLE at the lowest SNR because the noise model was "perfect" for the application of the B-spline methodology, whereas WLE must numerically approximate the PDFs to have a particular signal and hence is at a disadvantage for the lowest signal to noise ratio as the higher amount of noise and limited amount of data most likely led to a higher error rate in approximating the PDFs. Though, the improvement was minimal

at best. The results of this experiment suggest that even when the PSF is optimized for the B-spline methodology, WLE is approximately equivalent.

Chapter 4

A Pairwise Distance Distribution Correction (DDC) algorithm to eliminate blinking-artifacts in super-resolution microscopy

[1]

4.1 Introduction

In recent years the development of superresolution fluorescence microscopy has enabled the probing of macromolecular assemblies in cells with nanometer resolutions. Amongst different superresolution imaging techniques, single-molecule localization superresolution microscopy (SMLM) has gained wide popularity due to its relatively simple implementation, which is based on post-imaging analysis of single-molecule detection.

[1]Bohrer CH, Yang X, Weng X, Tenner B, Ross B, Mcquillen R, Zhang J, Roberts E, Xiao J. A Pairwise Distance Distribution Correction (DDC) algorithm for blinking-free super-resolution microscopy. BioRxiv. 2019 Jan 1:768051

SMLM reconstructs a superresolution image by stochastic photo-activation of individual fluorophores and subsequent accurate post-imaging localization determination [5, 6, 7]. One major advantage of SMLM is that due to its single-molecule detection nature, one can determine the number of molecules in a macromolecular assembly quantitatively, allowing the investigation of both the molecular composition and spatial arrangement at a level unmatched by other ensemble imaging-based superresolution imaging techniques. In the past few years SMLM has led to novel discoveries and quantitative characterizations of numerous biological assemblies [161, 162] such as those composed of RNA polymerase [163, 164, 165], membrane proteins [166], bacterial divisome proteins [167, 168, 169, 139], synaptic proteins [170, 171], the cytoskeleton [172], DNA binding proteins [88, 173], chromosomal DNA [174], viral proteins [175], and more.

One critical aspect in realizing the full quantitative potential of SMLM relies on the careful handling of the blinking behavior of fluorophores. A photo-switchable fluorophore can switch multiple times between activated and dark states before it is permanently photobleached, leading to "repeat localizations" from the same molecule. These repeat localizations are often misidentified as multiple molecules adding an additional degree of noise to the superresolution images — resulting in the appearance of blinking-artifacts (usually in the form of false nanoclusters), an increase in error for almost any quantification, and counting errors (making it challenging to quantify numbers of molecules and

the stoichiometry of complexes) (Fig. 4.1A) [176, 177, 178, 179, 180].

Multiple groups have developed different methods to correct for blinking-artifacts within SMLM. These methods can be coarsely divided into two categories depending on whether a method provides a corrected image void of repeat localizations or a statistical analysis summarizing the properties of the image at the ensemble level. Methods in the first category commonly use a variety of threshold values both in time and space to group localizations that likely come from the same molecule [5, 6, 176, 181, 178, 180]. The advantage of using thresholds is that it results in a corrected image, allowing one to observe the spatial distribution of fluorophores in cells and apply other quantitative analyses as needed. The disadvantage is that a constant threshold value is often insufficient in capturing the stochastic nature of fluorophore blinking and heterogeneous molecular assemblies. Furthermore, calibration experiments and/or a priori knowledge of the fluorophore's photochemical properties are often needed to determine the appropriate threshold values [178, 176, 182, 19, 180]. Statistical analyses such as maximum likelihood or Bayesian approaches have been developed to take into account the stochastic behavior of blinking to count the number of fluorophores, but have yet to produce a corrected superresolution image void of repeat localizations [183, 184, 185]. Additionally, many of these approaches are dependent on specific photokinetic models for the fluorophore, which can be complex and difficult to determine [186, 20, 182, 187, 19, 188].

The second category of methods analyze raw, uncorrected SMLM images using statistical methods to characterize the mean properties of the organization

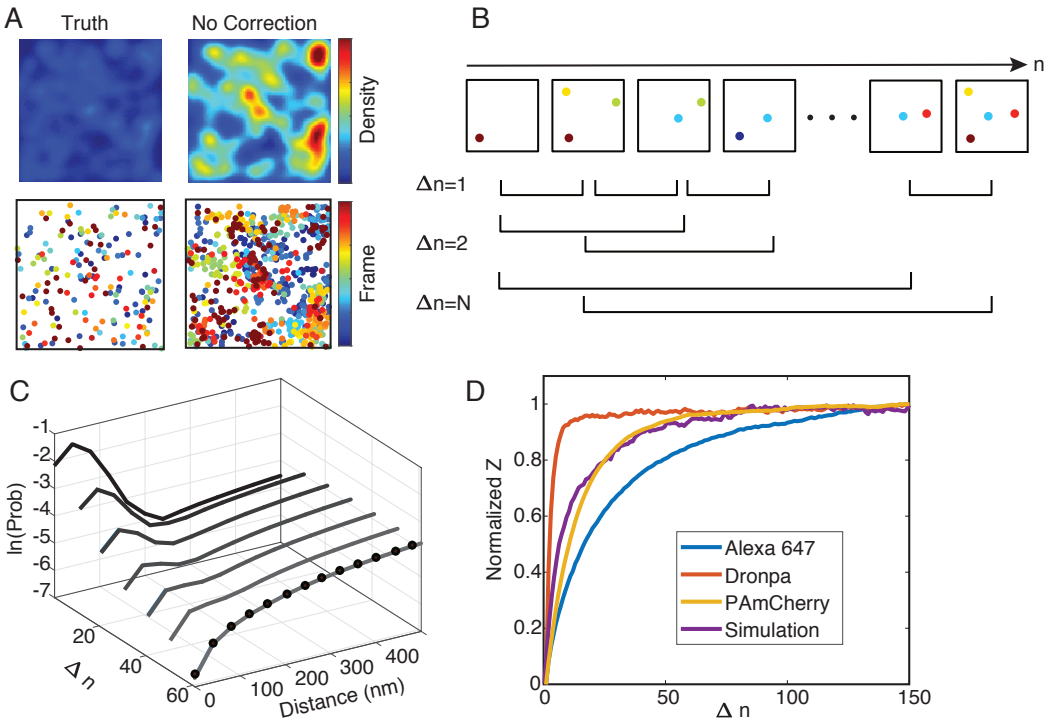

Figure 4.1: A. Simulated SMLM superresolution images (top panel) of randomly distributed molecules without repeats (Truth) and with repeats (No correction). The corresponding scatter plots (colored through time) are displayed in the bottom panel. B. Schematics of how the pairwise distance distributions at different frame differences (Δn) were calculated. C. Pairwise distance distributions at different Δn (black to gray curves) converge to the true pairwise distribution (black dots) when Δn is large. D. Normalized Z values measured for three commonly used fluorophores and a simulated fluorophore as that used in A. All Z values reach plateaus at large Δn, indicating that at large Δn, the pairwise distance distributions converge to a steady state. The normalized Z value was calculated by taking the difference between the cumulative pairwise distance distribution at a Δn and that at $\Delta n = 1$: $(Z(\Delta n) = \sum |cdf(P_d(\Delta r|\Delta n)) - cdf(P_d(\Delta r|\Delta n = 1))|)$.

of molecules at the ensemble level. Pair- or auto-correlation-based analyses (PCA) have been used extensively within the field [179, 189]. The long tail of the correlation function can often be fit to a specific model to extract quantitative parameters. This class of methods is prone to model-specific errors, especially if the underlying structures of the molecular assemblies are heterogeneous and vary throughout the image [144]. A recently developed method analyzes the clustering of a protein with experimentally varied labeling densities, which was robust in determining whether membrane proteins form nanoclusters and was insensitive to many imaging artifacts [177]. A post-imaging computational analysis capitalizing on the same principle has also been developed [190]. Although these methods are powerful in determining whether a protein of interest forms clusters or not, they provide a quantification at the ensemble level but not a corrected image, which limits their use in analyzing heterogeneously distributed molecular assemblies and their spatial arrangement in cells.

Here, we present an algorithm, termed Distance Distribution Correction (DDC), to enable robust reconstruction and quantification of blinking-artifact-free SMLM superresolution images without the need of setting empirical thresholds or performing experiments to calibrate a fluorophore's blinking kinetics. We first validate our approach using a diverse set of simulated and experimental data and compare DDC to other existing methods. In each situation DDC outperformed the existing methods in obtaining the closest representation of the "true" image, minimizing noise within various analyses, and in determining

the accurate number of fluorophores. We then applied DDC to experimentally collected SMLM images of two orthologs of a scaffolding protein that is important for the organization of membrane microdomains, A-Kinase Anchoring Protein 79/150 (AKAP79 and AKAP150) [191, 192, 193]. Both proteins showed clustered organizations, but with significantly reduced numbers and sizes of clusters when compared to the commonly used thresholding method, changing the quantitative properties of membrane microdomains organized by these proteins. Finally, we discuss critical considerations of how to apply DDC to experiments successfully.

4.2 Results

4.2.1 Principle of DDC

DDC is based on the principle that the pairwise distance (Δr) distribution, $P_d(\Delta r|\Delta n)$, of the localizations separated by a frame difference (Δn) much larger than the average number of frames a molecule's fluorescence lasts (N) approximates the true pairwise distance distribution $P_T(\Delta r)$. Note that N does not need to be precisely determined as long as it is in the regime where $P_d(\Delta r|\Delta n)$ approaches a steady state, as we show below. One intuitive way to understand this principle is that, if one collects an imaging stream that is long enough so that all the localizations in the first and last frames of the stream come from distinct sets of fluorophores, the pairwise distance distribution between the localizations of the two frames will then be devoid of blinking and

will reflect the true pairwise distance distribution ($P_T(\Delta r)$). A mathematical justification of this principle is provided in the supplemental material with an in-depth discussion and illustration (Fig. 4.2).

To demonstrate the principle of DDC, we used simulated SMLM images of randomly distributed fluorophores that followed the photokinetic model shown in Fig.4.3A. One representative superresolution image and the corresponding scatter plot, colored through time, with and without repeat localizations are shown in Fig. 4.1A. Apparent clustering was observed in images when blinking-artifacts were not corrected. Using the uncorrected images, we computed the pairwise distance distributions at all frame differences Δn (Fig. 4.1B). As shown in Fig. 4.1C and Fig.4.4, at small Δn there are large peaks at short distances, indicating that there were repeat localizations from the same fluorophores closely spaced in time and space. When Δn is large, the pairwise distance distributions approach a steady state converging upon the true pairwise distance distribution (Fig. 4.1C, dotted curve). This behavior supports the principle that when Δn is large the pairwise distance distribution represents the true pairwise distance distribution. Using simulations, we also show that the pairwise distance distributions converge upon the true distributions at large Δn irrespective of the underlying photokinetics or molecular spatial distributions (Fig.4.4, Methods Section).

Next, we used experimentally obtained SMLM images of three molecular assemblies labeled with different fluorophores in *E. coli* cells, the bacterial tran-

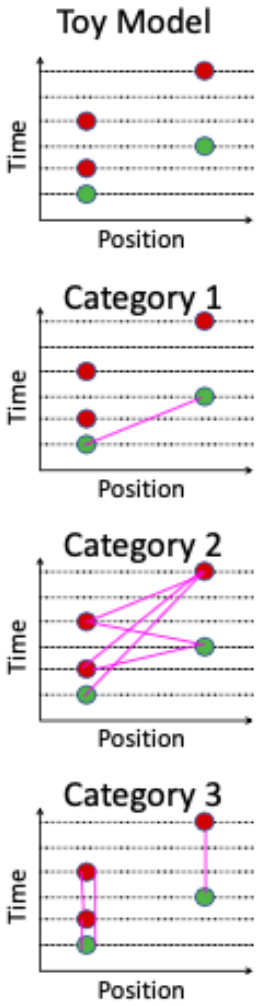

Figure 4.2: The top row shows a simple one dimensional system illustrating the blinking of two fluorophores, where the green dots are the true localizations and the red dots are repeats. The subsequent rows show the different categories referenced within the Methods Section, with the pink lines illustrating the pairs of localizations for each category.

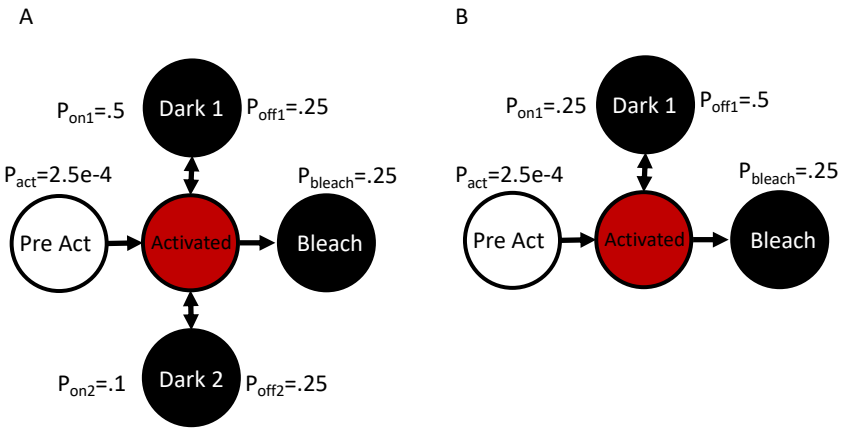

Figure 4.3: The two kinetic models used to simulate blinking, A.) 2 dark state and B.) 1 dark state. The transition probabilities per frame are shown in the figure.

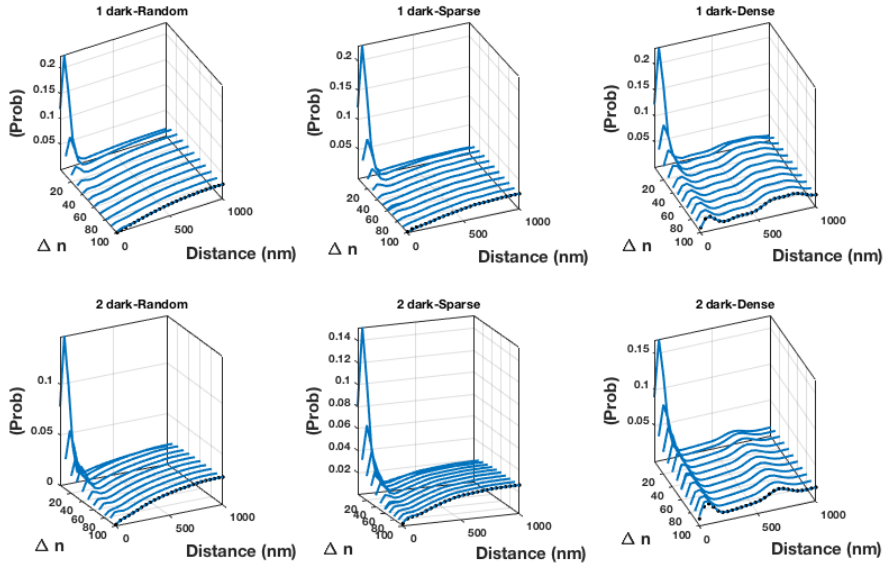

Figure 4.4: The pairwise distance distributions for both photo-kinetic models shown in Fig.4.3 and 6 molecular assemblies. Note here that the axis is no longer log scale as in the main text and the true pairwise distance distribution is shown as black dots.

scription elongation factor NusA fused with the reversibly switching green fluorescent protein Dronpa [194], *E. coli* RNA Polymerase fused with the photoactivatable red fluorescent protein PAmCherry [34], and precursor ribosomal RNAs (pre-rRNA) labeled with organic fluorophore Alexa647-conjugated DNA probes [195] (Fig. 4.5, Methods Section). We determined the pairwise distance distribution for each fluorophore and calculated the normalized, summed differences of the cumulative distributions for each Δn, relative to that of $\Delta n = 1$, ($Z(\Delta n) = \sum |cdf(P_d(\Delta r|\Delta n)) - cdf(P_d(\Delta r|\Delta n = 1))|$). As shown in Fig. 4.1D, in all cases the corresponding normalized Z reach plateaus at large Δn despite different photokinetics and spatial distributions. The rate at which each fluorophore reaches the plateau for the normalized Z reflects the photokinetics of the fluorophore — the longer a fluorophore blinked (such as Alexa647 compared to Dronpa), the longer the time until Z plateaued. These experimental results further verify the principle of DDC by showing that the pairwise distance distributions converge upon a steady state distribution as Δn increases.

It is important to note that the determination of $P_T(\Delta r)$ is not dependent upon a particular photokinetic model of the fluorophore nor does it require experimental characterizations of the fluorophore. $P_T(\Delta r)$ can be determined solely from the SMLM image stream as long as it is long enough so that a steady state of $P_d(\Delta r|\Delta n)$ can be reached (Fig. 4.1C, Fig.4.4).

Once determined, $P_T(\Delta r)$ can then be used to calculate the likelihood to have

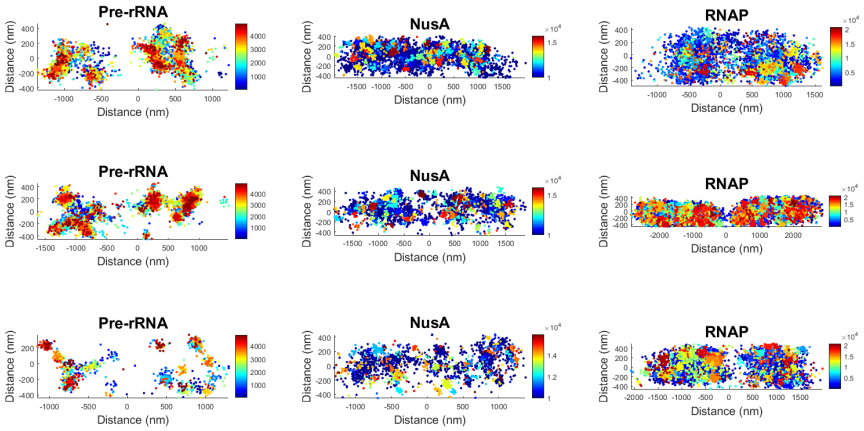

Figure 4.5: Example scatter plots of the experimental data used to verify that the pairwise distance distributions reached a steady state distribution. We show 3 cells for each molecular assembly, with the localizations colored with the frame of the localization.

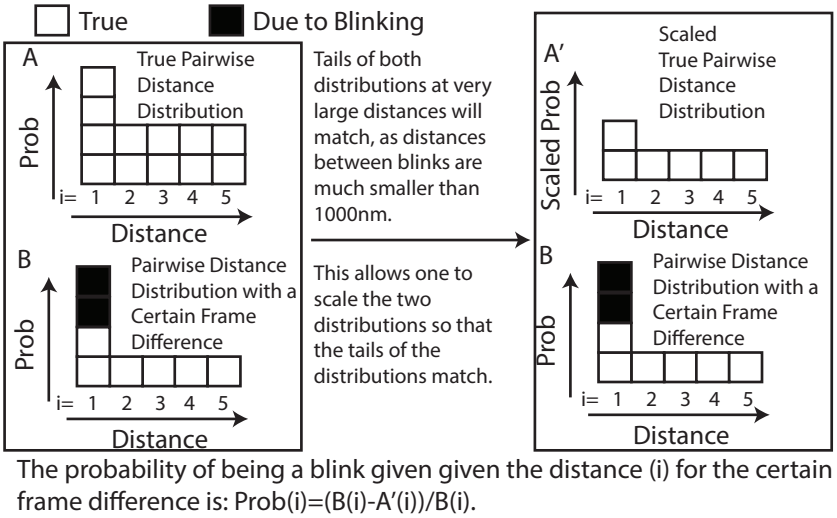

Figure 4.6: An illustration showing how to calculate $\mathbf{M}$ using the pairwise distance distributions. The blocks represent the distributions and i is the distance bin.

a particular subset of true localizations (Fig. 4.6-4.9, Methods Section) using the following equation:

$$\mathcal{L}(\{R,T\}|\mathbf{r},\mathbf{n}) = \prod_{i,j\in\{T\}} P_T(\Delta r_{i,j}) \times \prod_{i\in\{R\},j\in\{R,T\}} P_{R1}(\Delta r_{i,j}|\Delta n_{i,j}), \qquad (4.1)$$

where $\{R,T\}$ are sets that contain the indices of the localizations that are considered repeats $\{R\}$ and the true localizations $\{T\}$ given the coordinates $\mathbf{r}$ and associated frame numbers $\mathbf{n}$ obtained from experiment. The first term on the right of the equation is the probability of observing all distances Δr between every pair of true localizations (i & $j \in \{T\}$). Here the probability distribution $P_T(\Delta r_{i,j})$ is the true pairwise distance distribution. The second term is the probability of observing all distances between pairs of localizations with at least one being a repeat ($i \in \{\mathrm{R}\}$ and $j \in \{R,T\}$). Here, the probability distribution $P_{R1}(\Delta r_{i,j}|\Delta n_{i,j})$ gives the probability of observing a distance between a pair of localizations with a frame difference $\Delta n_{i,j}$ if at least one of the localizations is a repeat. This probability distribution can be easily determined once $P_T(\Delta r)$ is known (Methods Section). Here, maximizing the likelihood with respect to $\{R,T\}$ results in a subset of true localizations where the pairwise distance distributions $P_d(\Delta r|\Delta n)$ are equal to $P_T(\Delta r)$ (Fig. 4.7). DDC maximizes the likelihood with respect to the two sets ($\{\mathrm{R},\ \mathrm{T}\}$) using a Markov Chain Monte Carlo (MCMC) [9, 8], to result in the corrected image (Fig. 4.8 and 4.9, Methods Section).

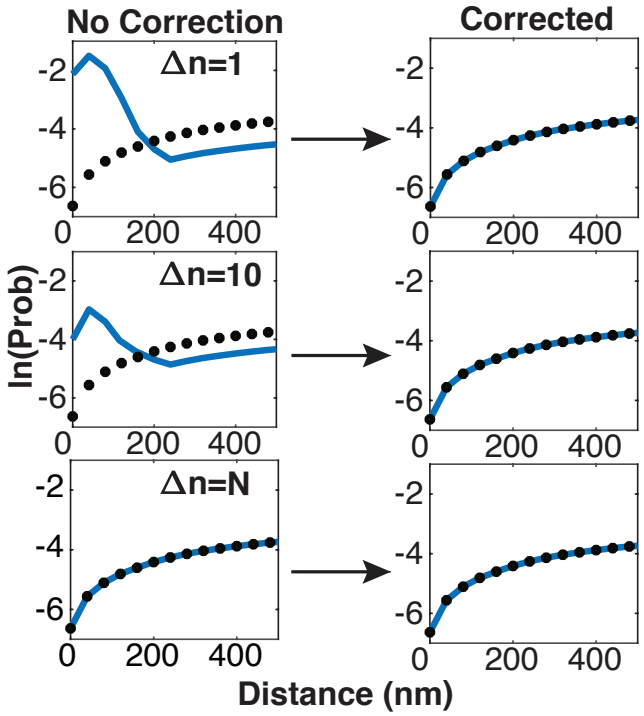

Figure 4.7: An illustration of the pairwise distance distributions at a certain frame difference, Δn, before and after being corrected with DDC. When the likelihood is maximized all of the pairwise distance distributions will match the true pairwise distance distribution. [The true pairwise distance distribution is shown as black dots.]

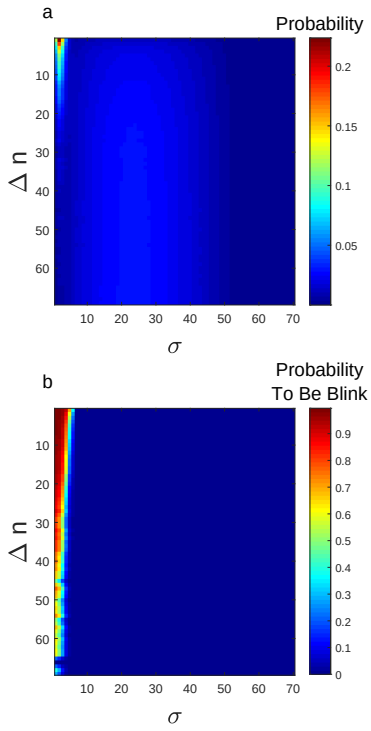

Figure 4.8: a. The probability distribution to observe a distance for a given Δn , in units of resolution σ, between two localizations when at least one of them is a repeat, $P_{R1}(\Delta r|\Delta n)$. This specific distribution is for the 1 dark state no clusters system. (See Methods Section text for details as to how these distributions are used to calculate Likelihood) b. The probability that a localization is the repeat of a given localization given the frame and distance between the localizations. These probabilities are calculated using the calculation shown in the prior figure.

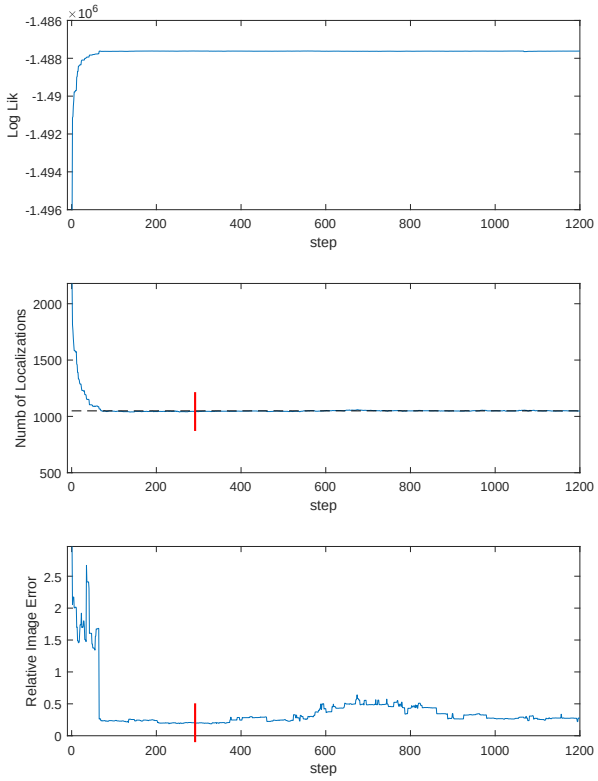

Figure 4.9: An example of the MCMC phase space search for the 2 dark state Small clusters system. For the number of localizations subplot a dashed black line shows the true number of localizations. For the bottom two subplots we show red lines indicating where the Likelihood was maximized. [Note: here we chose a random starting position for $\kappa(density)$ to illustrate the burn in phase of the MCMC, when $\kappa(density)$ starts at zero the burn in phase is not so extreme.]

To validate Equation 1, we show that only when greater than 97% of the final localizations are the true localizations does the likelihood reach its maximum (Fig. 4.10). This result was observed regardless of distinct spatial distribution or photo-kinetics of the fluorophore in six different simulations (Fig. 4.10).

4.2.2 DDC outperforms existing methods in both image reconstruction and counting the number of molecules

To compare the performance of DDC with commonly used thresholding methods, we simulated four systems, random distribution (no clustering), small clusters, dense clusters, and filamentous structures (Fig. 4.11, Methods Section). In these simulations the fluorophore had two dark states and followed the photokinetic model shown in Fig.4.3A. The raw images without any blinking-artifact correction for each simulation are shown in Fig. 2A. We applied DDC, three published thresholding methods (T1 to T3 [176, 180, 178])(Methods Section, Fig. 4.12 and 4.13) and a customized thresholding method (T4, Methods Section) to all the images. Method T1 links together localizations using a time threshold that is determined using an empirical estimation of the photokinetics of the fluorophore [176] (Fig. 4.12, Methods Section). Method T2 uses the experimentally quantified photo-kinetics of the fluorophore to set extreme thresholds so that the possibility of overcounting is extremely low [180]. Method T3 uses the experimentally determined number of repeats per fluorophore to choose thresholds that result in the correct number of local-

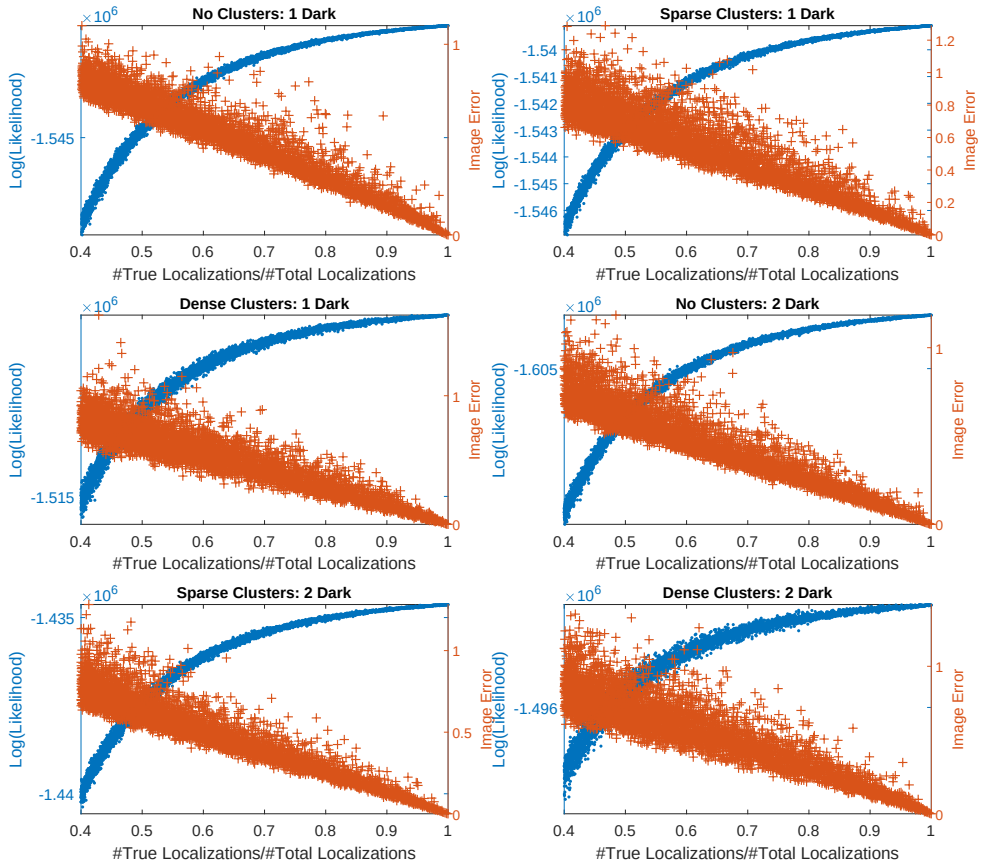

Figure 4.10: Maximization of Likelihood Results in Correct Conformation of Localizations: For 6 systems investigated within this work, we randomly varied the percentage of true localizations and calculated the $log(Lik)$ and the image error for each conformation.

izations within each image [178](Fig. 4.13, Methods Section). T2 and T3, but not T1, require additional experiments to characterize fluorophore photo properties. Method T4 is a customized, ideal thresholding method that scans all possible thresholds and uses the thresholds that result in the least Image Error for each system (Methods Section). T4 cannot be applied in real experiments since the true, blinking-artifact-free image is unknown — we included it here to illustrate the best scenario of what a thresholding method could achieve. To quantitatively compare the ability of these methods in producing a blinking-artifact corrected image we calculated two metrics, the Image Error and Counting Error (Fig. 4.11B, Methods Section). The Image Error was calculated by first summing the squared difference of each pixel's normalized intensity between the corrected images and the true image, and then dividing this squared difference by the error between the uncorrected image and the true image (Methods Section). The Image Error quantifies the amount of error in determining the distribution of localizations without being penalized for the error in the number of localizations. The Counting Error was calculated as the difference between the true number of fluorophores and that determined from the corrected image divided by the actual number of fluorophores (Methods Section).

As shown in Fig. 4.11B, DDC outperforms all four methods by having the lowest Image Errors and lowest (or close-to-lowest) Counting Errors. Interestingly, even with the best possible thresholds (T4), DDC still outperforms T4 in determining the correct spatial distribution and numbers of localizations. This

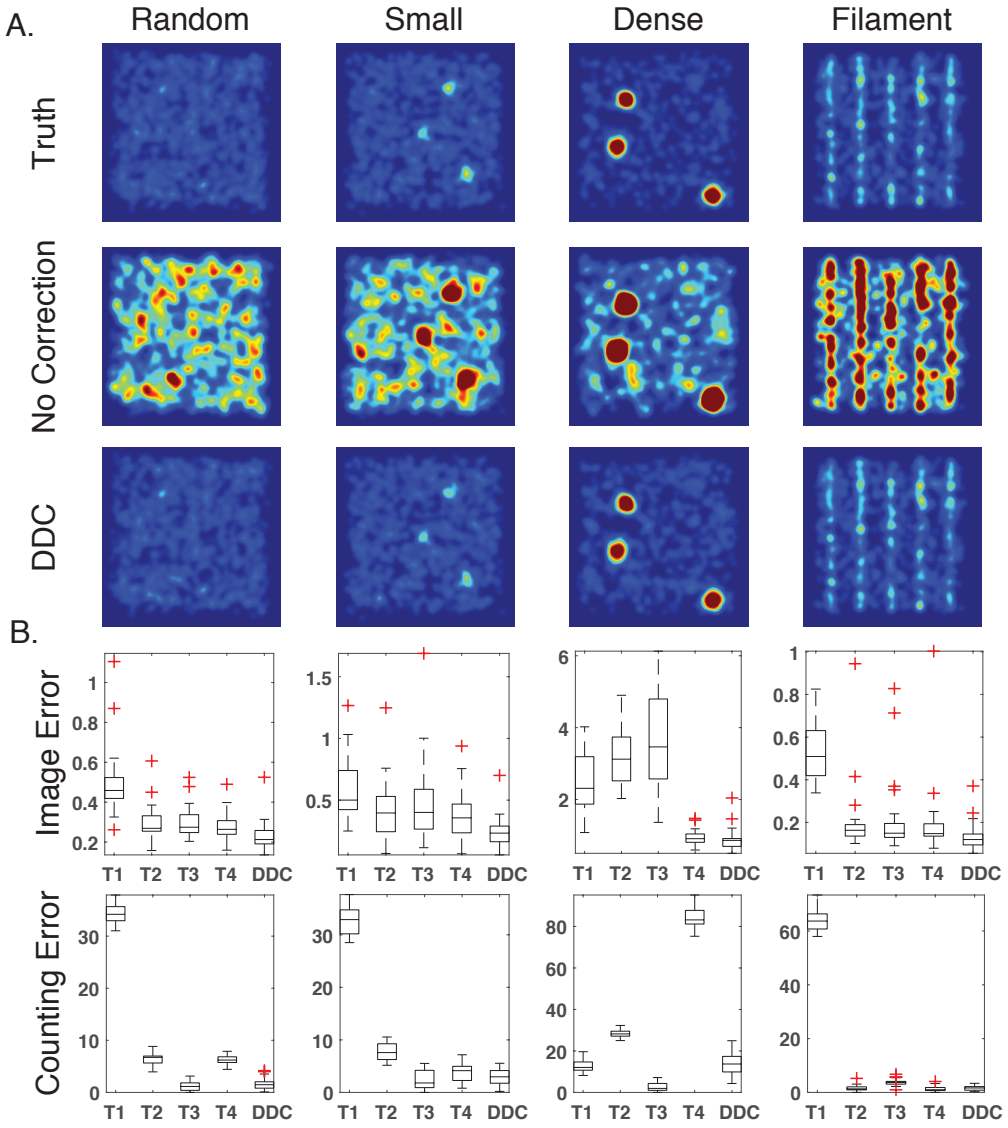

Figure 4.11: Comparison of four different thresholding methods with DDC on four spatial distributions (randomly distributed, small clusters, dense clusters and filaments). A. True, uncorrected and DDC-corrected images for each spatial distribution. B. Image Error and Counting Error calculated from T1 to T4 and DDC for each spatial distribution. The whiskers extend to the most extreme data points not considered outliers, and the red pluses are the outliers (greater than 2.7 std).

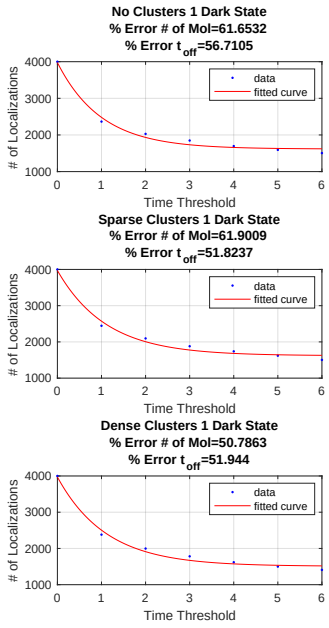

Figure 4.12: Resulting Error in Using Methodology of Annibale et al. (1): Here we only show the results for the 1 dark state systems with the fits to the semi-empirical formula (See Text). In the titles of each subplot we show the percent error in determining the number of true localizations and the average dark time.

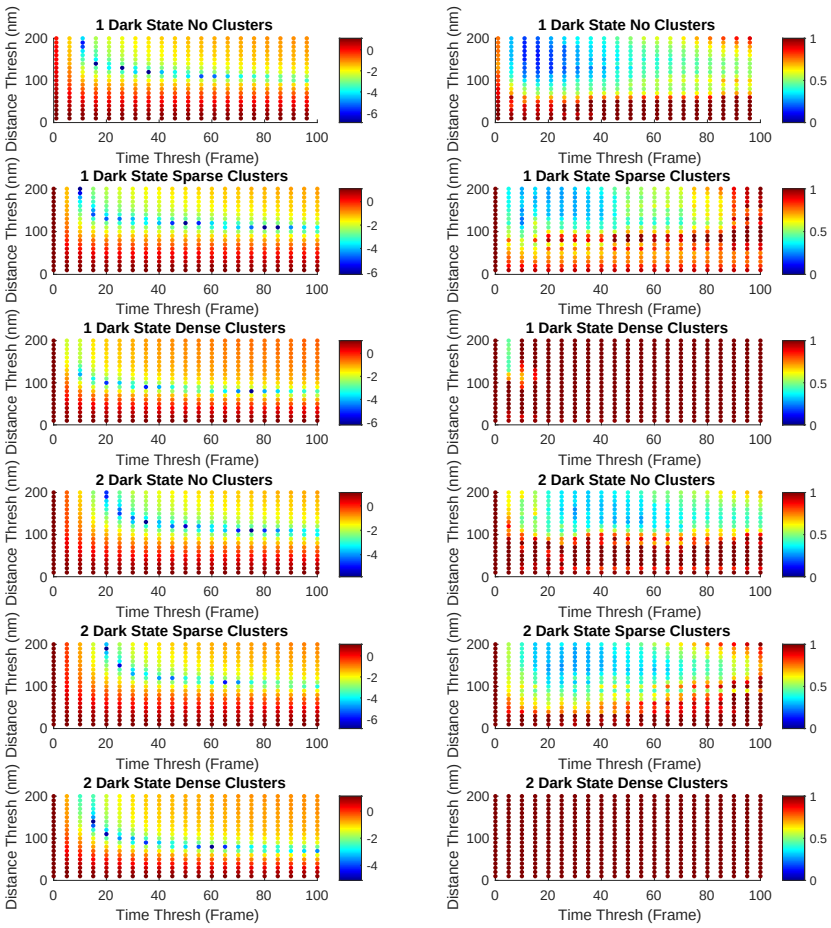

Figure 4.13: Determining the Thresholds for the Coltharp et al. Approach: In the first column we show the difference from the true number of localizations for the various time thresholds and distance thresholds, log scale ($ln[abs(\#loc - \#loc_{true})/\#loc_{true}]$. In the second column we plot the Image Error for each pair of threshold values for six systems.

result suggests that thresholds cannot adequately account for the stochastic nature of blinking. Similar results are shown in Fig. 4.14 for a fluorophore with one dark state (Fig.4.3B). When counting the number of localizations is the main concern, T3 performs equally or slightly better than DDC because T3 was applied with an experimental calibration that provides the average number of blinks per fluorophore (Fig. 2, Methods Section). Nonetheless, DDC outperforms T3 by having lower Image Errors across all four different simulation systems, especially for the dense cluster system, where the average Image Error of T3 is seven times that of DDC (Fig. 4.11B). In conclusion, these results indicate that DDC can be used to obtain the correct number of true localizations and at the same time produce the most accurate SMLM images.

4.2.3 DDC decreases noise in the quantification of sister chromatids and dynein motor proteins

Quantifications of molecular assemblies with superresolution techniques are becoming vital to our understanding of biology, and therefore the error within these measurements must be minimized. Repeat localizations add an additional source of noise to SMLM images, leading to error in the quantifications of molecular structures and the interpretation of said images. In this section, we examine the utility of DDC in regards to these concerns by showing how the variation in signal between sister chromatids differs for the different methodologies and then how the locations of dynein motor assemblies vary with the different blinking-artifact correction methodologies.

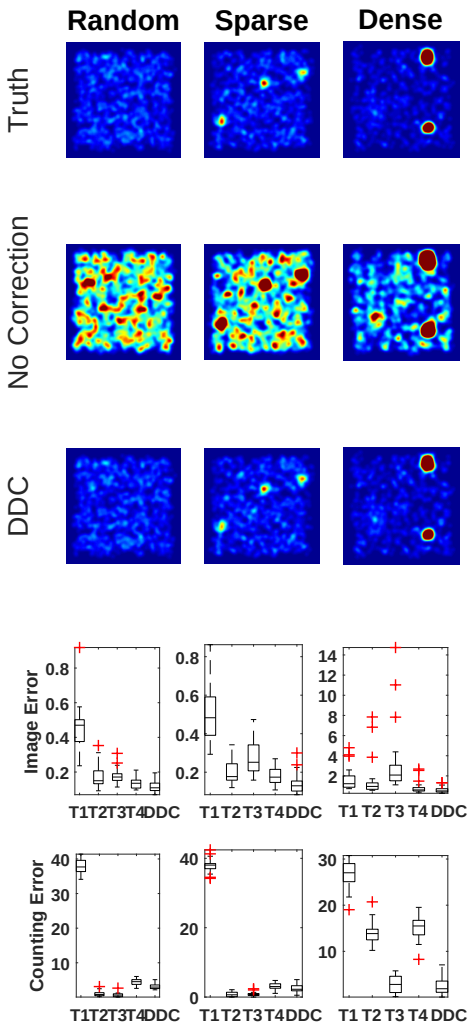

Figure 4.14: A comparison of the various thresholding methodologies with DDC and no blinking correction for the 1 dark state fluorophore. The first three rows show the images set to the same contrast for each labeled method. The last two rows show the results for the Image Error and the percent error in the number of fluorophores for each of the three systems for the one dark state fluorophore.

A powerful technique that has led to a deeper understanding of how stem cells differentiate asymmetrically is an adapted chromatin fiber technique [196]. The technique allows one to visualize sister chromatids and their associated proteins outside of the nucleus, providing a direct comparison between the two fibers. By quantifying the difference of old histone H4 to newly synthesized H4 between sister chromatids Wooten *et al.* showed that histone H4 is inherited asymmetrically in Drosophila melanogaster male germline stem cells during asymmetric cell division.

To investigate how blinking-artifacts can influence these measurements, we isolated sister chromatid fibers from Drosophilia embryos labeled with yoyo and performed SMLM imaging (Methods Section, Fig. 4.15A). We found that the sister chromatids exhibited a relatively regular structure when compared to the filament structure simulations of Fig. 4.11A — due to the stochasticity in the "labeling" of the simulation filaments (Methods Section). To quantify the performance of DDC on a more "regular" overlapping filamentous structure, we simulated overlapping filaments with no labeling variability and applied the various methodologies (Fig. 4.16). As expected, we found that DDC outperformed all other methodologies in regards to the Image Error and Counting Error (Fig. 4.16). We also observed for the first time that the T1 thresholding methodology outperformed T2 and T3 in terms of the Image Error, suggesting that the performance of the different thresholding methodologies is dependent upon the organization of the molecular structures.

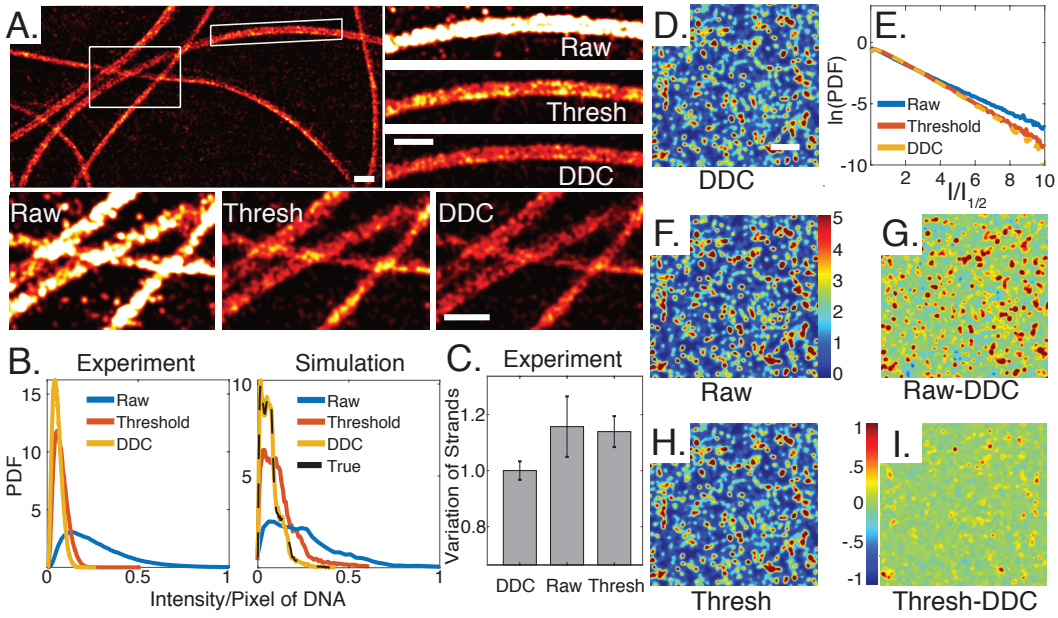

Figure 4.15: DDC minimizes the likelihood of misinterpreting images and analyses. A. SMLM images of sister chromatids with zoomed-in regions showing the images that result for each methodology (Raw (No correction), Threshold (T1) and DDC). (Scale bars, $1\mu m$) B. The intensity distributions of each methodology for the DNA pixels and the continuous filament simulations (Fig. 4.16). C. The amount of noise between sister chromatids for the different methods (Fig. S14). [D, F, H]. The three methodologies' SMLM images of dynein motor proteins (normalized by their medians, $I_{1/2}$). E. The methodologies' intensity distributions normalized by their medians. [G and I]. The difference between the methodologies' SMLM images with DDC's SMLM images.

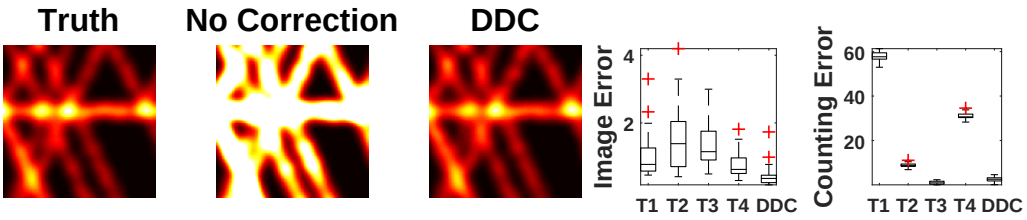

Figure 4.16: The comparison of the four different thresholding methodologies with DDC on the regular overlapping filamentous simulation system. The fluorophore for these simulations was that of Fig. 4.3A with a localization precision of 20nm. For the regular overlapping filamentous simulation system, there was zero noise in labeling density.

We then compared the images of the sister chromatids with no correction (Raw), T1 (Thresh) and DDC and saw major differences in the density of localizations along the filaments (Fig. 4.15A, insets). The extent of signal variability throughout the DNA fibers for the different methodologies (intensity per pixel of DNA) is shown in Fig. 4.15B for both the experimental data and simulation data, as well as the actual signal variability for the true localizations of the simulation data. We observed that the uncorrected localizations had the largest degree of variability with a long tail, then T1 and then DDC. Interestingly, we also found that the difference between T1 and DDC for the experiment was not as extreme as in the simulation — again suggesting the effectiveness of the thresholding methodology is specific for each system.

Next, we quantified a new metric, the Variation of Strands, which quantifies the noise in signal between sister chromatids and quantifies the amount of uncertainty in comparing the signal between sister chromatids — a measurement

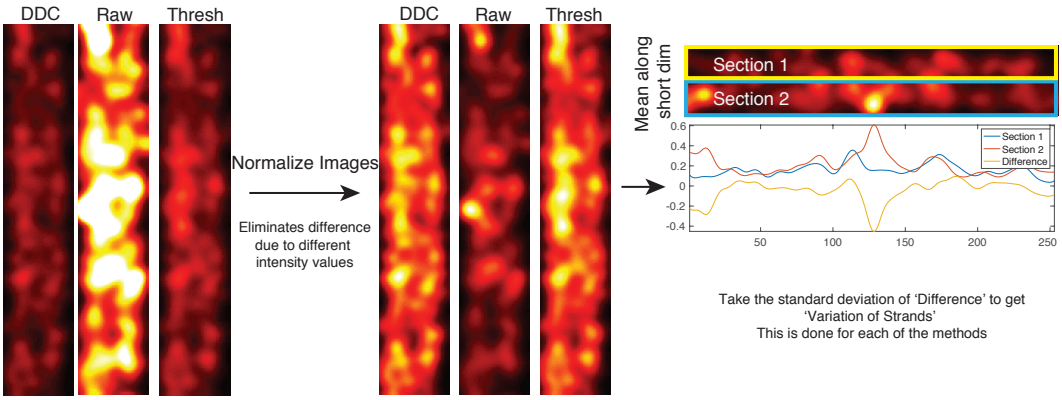

Figure 4.17: An illustration showing the methodology for the calculation of the Variation of Strands. For the normalization step the local image of the two sister chromatids signals are normalized to have a range between 0 and 1.

used in Wooten *et al.* [196]. The Variation of Strands was calculated by taking the standard deviation of the normalized difference in signal between the sister chromatids (Fig. 4.17). We quantified the Variation of Strands for the no correction (Raw), T1 (Thresh) and DDC methodologies (Fig. 4.15C) and observed that the Variation of Strands (relative to that from DDC) was significantly higher for both the raw and threshold methodologies when compared to that of DDC. As expected, this suggests that not properly correcting for repeat localizations does significantly affect the accuracy of this type of analyses.

These results are interesting, as the signal variabilities of the thresholding methodology and DDC for the experimental data were similar (Fig. 4.15C), and yet the thresholding and no correction methodologies produced similar Variations of Strands. These experimental results are similar to the results of the simulation systems (Fig. 4.11) — where a low Counting Error (similar to

signal variability Fig. 4.15B) does not necessarily indicate a low Image Error (similar to Variation of Strands) (Fig. 4.15C).

To experimentally investigate this further, we performed SMLM imaging of dynein within HeLa IC74 cells and quantified the differences in the locations of dynein signal with the different methodologies (Methods Section). Previous research suggests that dynein forms oligomers, and therefore dynein provides an excellent experimental system to investigate how blinking-artifacts influence the locations of the various dynein assemblies [197]. After applying the methodologies, we normalized the images by their medians (Fig. 4.15D,F,H) ($I/I_{1/2}$) so each method's image had an comparable signal distribution (Fig. 4.15E, note the log scale). Assuming the signal ($I/I_{1/2}$) was directly proportional to the number of dynein in a pixel for each methodology, we calculated the difference of the normalized images for the no correction methodology (Fig. 4.15F) and the thresholding methodology (Fig. 4.15H) with the normalized image from DDC (Fig. 4.15D). As shown in Fig. 4.15G and Fig. 4.15I in certain regions of the images the signal is greater and lower in others (relative to DDC), indicating that the spatial distribution of signal for the different methodologies can be very different, suggesting that if blinking-artifacts are not properly corrected, a combination of Image Error and Counting Error will lead to misinterpretations as to where the different molecular assemblies reside (in this case, the oligomerization states of dynein).

4.2.4 DDC identifies differential clustering properties of membrane microdomain proteins AKAP79 and AKAP150

Membrane microdomains formed by membrane proteins have been commonly observed in super-resolution imaging studies and have raised significant interest in their molecular compositions and associated biological functions [166]. However, concerns remain as of whether the characterizations of these microdomain protein clusters were impacted by blinking-artifacts [177]. Here we used DDC to investigate a membrane scaffolding protein, A-Kinase Anchoring Protein (AKAP), which plays an important role in the formation of membrane microdomains [191, 192, 193]. The two orthologs AKAP79 (human) and AKAP150 (rodent) were previously shown to form dense membrane clusters, which are likely important for regulating anchored kinase signaling.

We performed SMLM imaging on AKAP150 in murine pancreatic beta cells using an anti-AKAP150 antibody and analyzed the resulting SMLM data using DDC (Methods Section). For AKAP79, we applied DDC to previously acquired SMLM data from HeLa cells [191]. For comparison, we also applied the T1 method to both scaffolding proteins as it was used in the previous study of the AKAP79 [191, 176] (Fig. 4.18, 4.19). We found that the images from DDC still showed significant deviations from what was expected from simulated random distributions, indicating the presence of clustering. We then verified this result using a new statistical analysis that summarizes the properties of the image at the ensemble level and has been shown to be

robust in its ability to differentiate random from clustered distributions of molecules (Fig. 4.20) [190]. We also observed that DDC images exhibited dramatically reduced clustering when compared to the uncorrected and T1-corrected images for both proteins (Fig. 4.21A). To quantitatively compare these images, we used a tree-clustering algorithm (Methods Section) to group localizations in individual clusters and show the corresponding cumulative distributions in Fig. 4.21B. The cumulative distributions show that the degrees of clustering for both proteins are significantly reduced when DDC was applied. Interestingly, AKAP150 shows a higher degree of clustering when compared to AKAP79, with more than 50% of the localizations within clusters containing greater than 15 localizations, twice that of AKAP79. Nevertheless, DDC-corrected AKAP79/150 images show significant deviations from the simulated random distributions, indicating the presence of clustering (Fig. 4.21B, compare yellow and purple curves). These results suggest that the clustering of the AKAP scaffolds are differentially regulated and the context dependence is likely important in considering the microdomain-specific signaling functions of the clusters.

4.2.5 Considerations in the application of DDC

As with any method, successful application of DDC to SMLM images requires an understanding of critical factors that could influence the performance of DDC. In this section, we evaluate the impact of localization density and activation rate on the performance of DDC using simulations. We also demonstrate that the commonly used practice of ramping the UV activation power

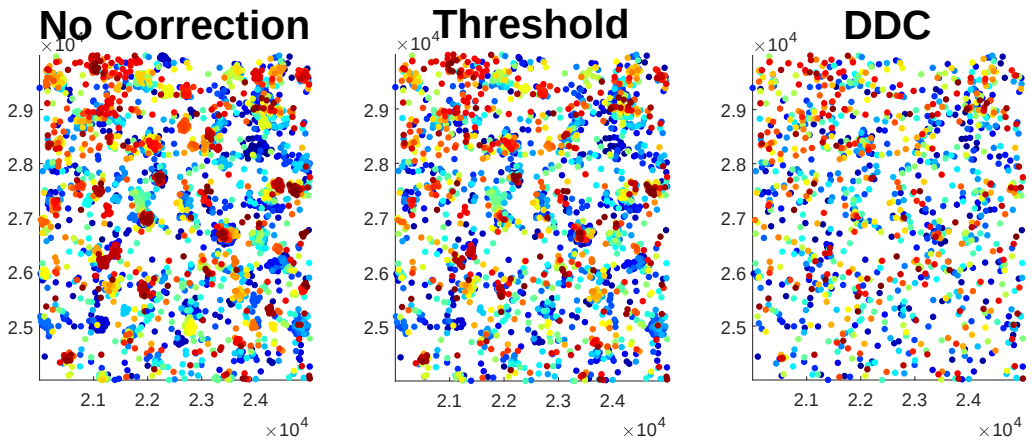

Figure 4.18: Scatter plots for a section of a cell with the localizations from AKAP79 with the color indicating the frame of the localization (Blue is early and Red is late). Here we show three different methodologies with the same thresholds used previously [191].

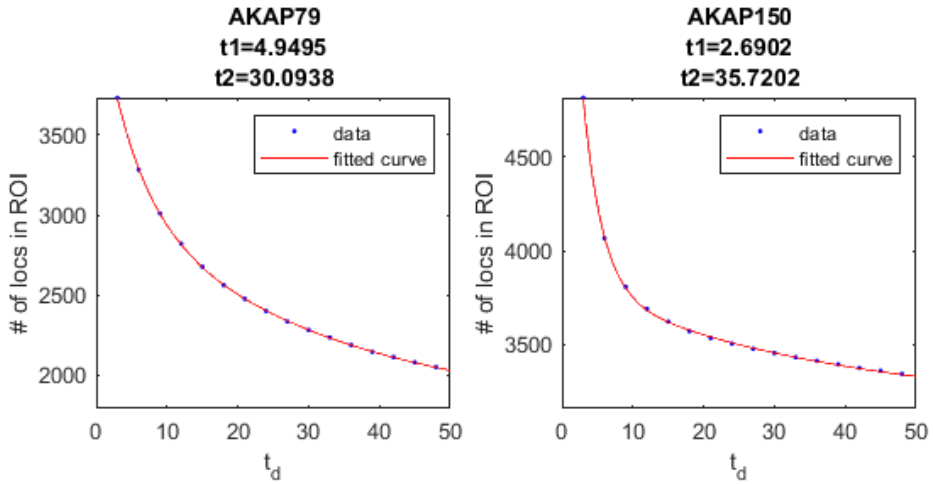

Figure 4.19: Here we show the results for determining the proper thresholds utilizing the methodology of T1 for AKAP79/AKAP150. The data was fitted to the double exponential used previously. Here the proper threshold is equal to two times the larger average dark time, either t1 or t2.

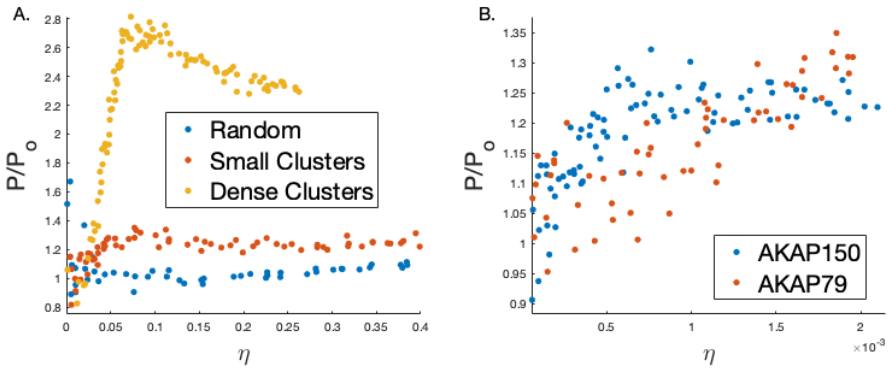

Figure 4.20: A. The results of computationally varying the label density on some of the simulation systems. B. The results of computationally varying the label density on AKAP79 and AKA150. (Values greater than 1 indicate significant clustering.)

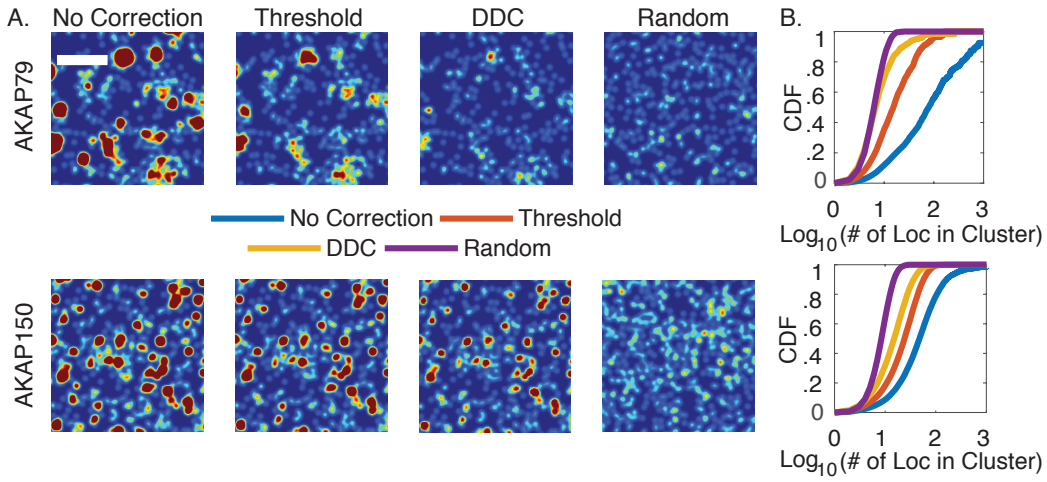

Figure 4.21: Application of DDC to experimentally measured spatial distributions of AKAP79 and AKAP150. A. SMLM images of the two scaffold proteins without correction, corrected using the thresholding method T1 and DDC, and that of a simulated random distribution using the same number of localizations of DDC-corrected images. B. Cumulative distributions for the number of localizations within each cluster for each protein. (Scale bar, $1\mu m$)

in SMLM imaging should be avoided when applying DDC.

To quantify the influence of localization density on the performance of DDC, we simulated random distributions of fluorophores with different densities ranging from 1000 raw localizations to 15000 localizations per $1\mu m^2$. Note that a density greater than 5000 localizations/μm^2 corresponds to a Nyquist resolution of 30 nm or better. As shown in Fig. 4.23A, the Image Error increases as the localization density increases and reaches a plateau at $\sim$ .35. We found that the increase in Image Error at high localization densities was mostly due to the decreased raw Image Error of the uncorrected images at high localization densities (Fig. 4.22A). The decreasing improvement of DDC at increasing sampling rate suggests that a high sampling rate of the underlying structure reduces the image distortion caused by repeats, although very high labeling densities ($>$ 10,000 localizations/μm^2) is usually difficult to achieve for protein assemblies.

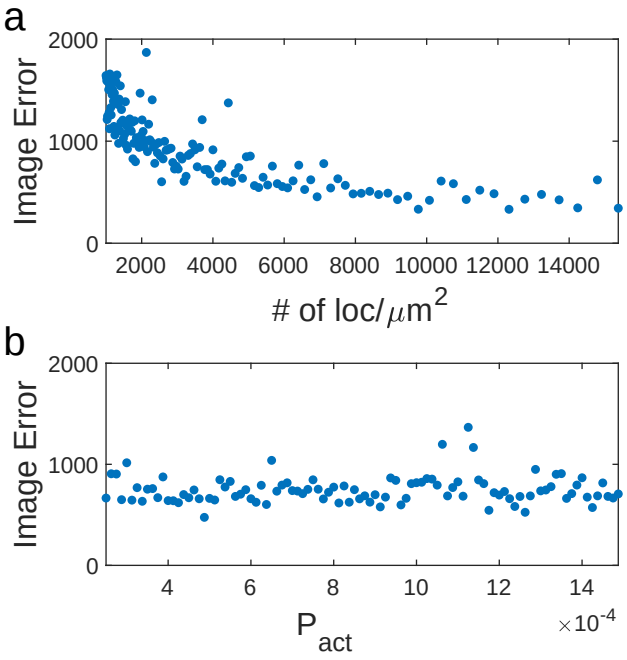

Figure 4.22: Here we show the raw Image Error (Not Normalized) for the uncorrected SMLM images for varying the density of the localizations and the activation energy.

Next, to quantify the influence of the activation rate, we varied the activation probability of each simulated fluorophore from .025 to .15 per frame, with 1000 fluorophores randomly distributed throughout a $1\mu m^2$ area. Fig. 4.23B shows that the Image Error of DDC steadily increases with the activation rate. This increase was because at high activation rates, the temporal overlaps of individual fluorophores that were spatially close to each other increased, which made it difficult to distinguish blinks from different fluorophores. Thus, as with all the other blinking-artifact correction methodologies, DDC obtains the best images when the activation rate is slow.

Finally, we illustrate one critical requirement for the successful application of DDC, that is, the photokinetics (blinking behavior) of the fluorophore, must be kept constant throughout the acquisition of the SMLM imaging stream (Methods Section). Note that this requirement is also needed for all other blinking-artifact correction methods [176, 178, 180]. One common practice in SMLM imaging is to ramp the activation power gradually throughout the SMLM imaging sequence in order to speed up the acquisition at later times when the number of fluorophores in the view field gradually deplete. The assumption is that activation power only changes the activation rate of a fluorophore (i.e. the probability of a fluorophore being activated per frame), but not the photokinetics of its blinking behavior (i.e. number of blinks, dark time and fluorescence-on time). Such a scenario indeed was shown for the photoactivatable fluorescent protein Dendra [19], but there are also reports showing that the photokinetics of mEos2 and PAmCherry are sensitive to the

activation intensity [19, 182].

We further investigated the activation dependence of the blinking behaviors of two commonly used fluorophores for SMLM imaging, the photoactivatable fluorescent protein mEos3.2 and the organic fluorophore Alexa647 with different activation (405nm) intensities. We quantified three parameters, number of blinks, off-times (T_{off}) and on-times (T_{on}), and report the mean value for each parameter as a function of activation intensity (Fig. 4.23C) We define one blink event as one continuous emission event that could span multiple fluorescence on-frames, the number of blinks as the number of repeated emissions separated by dark frames from the same fluorophore, T_{off} as the time between each blink and T_{on} as the time that the fluorophore remained fluorescent at each blink-on event (Fig. 4.23C). We observed that both fluorophores had a similar dependence of T_{on} with UV intensity, where T_{on} initially increased and then decreased at higher UV intensities (Fig. 4.23D, top), suggesting that UV also participates in the fluorescence emission cycle of the fluorophores. Next, we found that T_{off} decreased non-linearly as the UV intensity increased for both fluorophores (Fig. 4.23D, middle). Finally, we observed that the average number of blinks for the Alexa647 molecule increased dramatically with UV intensity while that of mEos3.2 remained largely constant (Fig. 4.23D, bottom), suggesting a differential influence of UV in changing the photokinetics of different fluorophores. Thus, varying the activation intensity during the acquisition of a SMLM image can indeed change the blinking characteristics of the fluorophores, which would affect the performance of DDC. These results suggest

that changing the activation intensity should only be done when a quantitative approach is not needed, or the proper controls have been performed to show that the fluorophore is insensitive to variations in the activation intensity.

4.3 Discussion

In this work we provided a blinking-artifact correction methodology, DDC, that does not depend upon exact thresholds, additional experiments, or a specific photo-kinetic model of the fluorophore to obtain an accurate reconstruction and quantification of SMLM superresolution images. DDC works by determining a "ground truth" about the underlying organization of fluorophores, the true pairwise distance distribution. We verified by simulations and experiments that such a true pairwise distance distribution can be obtained by taking the distances between localizations that are separated by a frame difference much longer than the average lifetime of the fluorophore. Using the true pairwise distribution, the likelihood can be calculated, where upon maximization of the likelihood one obtains an accurate representation of the true underlying structure.

We compared the performance of DDC with four different thresholding methods using simulated data with various spatial distributions and on fluorophores with different photokinetic models. DDC outperformed these methods by providing the "best" corrected images as well as excellent estimates of the number

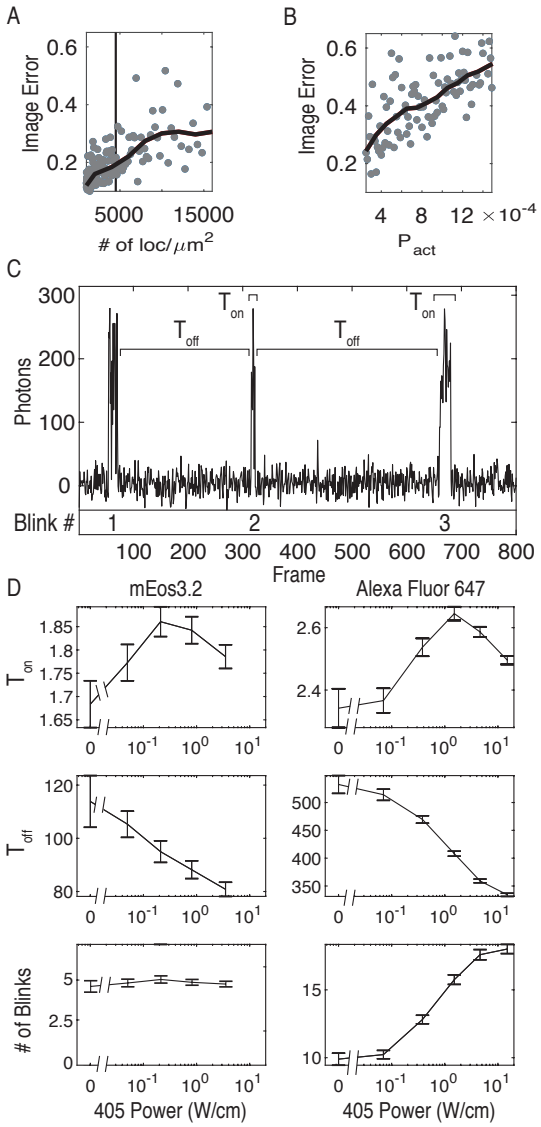

Figure 4.23: Image Error at different densities of localizations (A) and activation probability per frame (B). The raw data points are shown as gray points and the moving average is shown in black (Methods Section). C. An intensity trajectory of a single mEos3.2 molecule with labels showing the definitions of T_{on} and T_{off}. D. The average T_{on}, T_{off}, and number of blinks for Alexa647 and mEos3.2 at different UV activation intensities (405 Power, error bars are standard deviation of mean using two repeats).

of molecules in each image.

We then experimentally demonstrated that blinking-artifacts can add an additional degree of noise to quantitative analyses and can lead to misinterpretations of where different molecular assemblies reside. We also used DDC to investigate the spatial organizations of two scaffolding proteins AKAP79 and AKAP150, which have been shown to form microdomain-like structures [191, 193]. DDC resulted in significantly less degrees of clustering for the two proteins when compared to that resulted from the thresholding method. Most interestingly, DDC's ability to count the number of true localizations in SMLM images allowed quantitative comparison between the clusters formed by the two proteins: AKAP150 was about 2-fold more clustered than AKAP79. Such a difference in clustering could indicate that the two proteins are differentially regulated in separate cell types and this context dependence could be important for the signaling functions of the clusters. Further experiments are required to explore these possibilities. An additional note is that DDC only counts the number of emitters, which does not necessarily equal to the number of molecules that are labeled using dye-conjugated antibodies [197].

Finally, we demonstrated that the higher the activation rate and the density of fluorophores, the smaller the relative improvement of DDC. We also showed that in order to use DDC, the common practice of ramping the UV should be avoided in certain cases (depending upon the particular fluorophore), as we verified that mEos3.2 and Alexa647 exhibited activation power-dependent

photokinetics. In essence, DDC is best suited for SMLM imaging when quantitative characterizations of heterogenous cellular structures are required. The complete package of DDC is available for download at https://github.com/XiaoLabJHU/DDC.

4.4 Methods

4.4.1 Mathematical justification for true pairwise distance distribution

Here we provide a mathematical justification supporting the principle that the true pairwise distance distribution is obtained when the pairwise distances are taken between localizations separated by a frame difference much longer than the average lifetime of the fluorophore.

Blinking causes the position of a fluorophore to appear throughout multiple frames, we refer to the localizations from the same fluorophore as a blinking trajectory and we define the first localization in a blinking trajectory as the true localization and all subsequent localizations as repeats. An illustration of two blinking fluorophores for a one dimensional image is shown in Fig.4.2 with the true localizations of the fluorophores shown as green dots and repeats in red. For this justification we assume that the blinking behavior of the fluorophores are independent of each other and the photo-kinetics of the fluorophores are constant and uniform throughout the acquisition of the image. Note: this is one of the major assumptions needed to apply DDC.

The number of repeats for an arbitrary fluorophore, a, follows an unknown random variable, $num_b(a)$, and the determination of the true position of fluorophore a, x_a, is dependent upon the localization precision of the microscope, δ. For the toy model in Fig.4.2 we have no error in determining the position of the fluorophore for simplicity. The distances contributed by two arbitrary fluorophores within an image can then be split into the three arrays/categories below:

$$\left\{ \begin{array}{c} C1 = \sqrt{((x_a + \delta) - (x_b + \delta))^2} \\ C2(1:\gamma) = \sqrt{((x_a + \delta) - (x_b + \delta))^2} \\ C3(1:\gamma') \approx \sqrt{((\delta) - (\delta))^2} \end{array} \right\},$$

where $\gamma = (num_b(a) + num_b(b) + num_b(a) \times num_b(b))$ and $\gamma' = \sum_{n=0}^{num_b(a)} n + \sum_{n=0}^{num_b(b)} n$, are the number of distances contributed to the pairwise distance distribution for the different categories. Here we should note that the number of distances contributed by the repeats [C2 and C3] can be much higher than the distances contributed by the true localizations, $C1$. The pairs of localizations belonging to the three categories for the two fluorophores are shown in Fig.4.2 for reference.

The distances in each of the categories are separated in time by a certain number of frames, Δn. We define N as the maximum lifetime of a fluorophore. The fact that the fluorescent fluorophores have a limited lifetime creates con-

straints on the frame differences the distances in each category can posses. The possible frame differences for the distances in categories $C2$ and $C3$ are then the following:

$$\left\{ \begin{array}{c} \Delta n_{C1} - N < \Delta n_{C2} < \Delta n_{C1} + N \\ \Delta n_{C3} < N \end{array} \right\},$$

where Δn_{C1} is the frame difference between the true localizations in $C1$.

Notice that if we only use the distances between localizations that are separate in time by N, $\Delta n = N$, a pair of arbitrary fluorophores that have at least some localizations in their blinking trajectories with a frame difference of N will contribute a certain number of distances, from $C1$ and $C2$ and all of the distances in $C3$ will be eliminated.

Now, if we use the distances with $\Delta n = N$, the number of distances contributed from $C1$ and $C2$ from any pair of arbitrary fluorophores follows the unknown random variable ϕ. [The distances contributed by each pair of fluorophores follows the same unknown random variable because the photo-kinetics of each fluorophore is the same.]

Then, to obtain an accurate approximation of the true pairwise distance distribution, $P_T(\Delta r)$, we construct the probability distribution with a bin width δ, assume that the pairs of arbitrary fluorophores ($pairs(i)$) within each distance bin i is large, and use the distances between localizations that are separated in

time by N. The approximate true probability of observing a distance within bin i is then the following:

$$P_d^i(\Delta r|\Delta n = N) = \frac{\sum_{w=1}^{pairs(i)} \phi}{\sum_{q=1}^{All_{pairs}} \phi} \approx \frac{pairs(i) \times \overline{\phi}}{All_{pairs} \times \overline{\phi}} = \frac{pairs(i)}{All_{pairs}} = P_T^i(\Delta r), \quad (4.2)$$

where All_{pairs} is the number of pairs of fluorophores, $\overline{\phi}$ is the mean of the random variable and $P_d^i(\Delta r|\Delta n = N)$ is the bin i of the pairwise distance distribution between all localizations separated by the given frame. Equation 4.2 shows that with the previously mentioned assumptions the probability of finding a distance within each bin will be identical to that of the true pairwise distance distribution, justifying the principle. Note that each frame difference larger than N can be used to approximate the true pairwise distance distribution, therefore creating the pairwise distance distribution using all distances between localizations that are separated by a frame difference larger than N leads to an even better approximation of the true pairwise distance distribution.

4.4.2 The Inner Workings of DDC

4.4.2.1 Defining the Likelihood

Here we define the Likelihood as the following:

$$\mathcal{L}(\{R,T\}|\mathbf{r},\mathbf{n}) = \prod_{i,j\in\{T\}} P_T(\Delta r_{i,j}) \times \prod_{i\in\{R\},j\in\{R,T\}} P_{R1}(\Delta r_{i,j}|\Delta n_{i,j}), \qquad (4.3)$$

where $\{R,T\}$ are sets that contain the indices of the localizations that are considered the repeats $\{R\}$ and the true localizations $\{T\}$, where both sets account for every localization. The actual experimental data are stored within the two terms $\mathbf{r}$ & $\mathbf{n}$, with the prior containing the coordinates of every localization and the later containing the frame. The first term on the right determines the probability of observing all of the distances between every pair of true localizations. Here the probability distribution $P_T(\Delta r_{i,j})$ is the true pairwise distance distribution, which gives the probability of observing a distance Δr between the two localizations i & j if they are both true localizations. The second term is the probability of observing all of the distances between the pairs of localizations if at least one is considered a repeat. Here, the probability distribution $P_{R1}(\Delta r_{i,j}|\Delta n_{i,j})$ gives the probability of observing the distance between the pair of localizations given the frame difference between them if at least one of the localizations is a repeat. Note that every pair of localizations are within the likelihood calculation no matter which localizations are assigned to the sets $\{R$ & $T\}$.

Overall, by maximizing the Likelihood a subset of true localizations is determined, where the pairwise distances between the true localizations are independent of frame, Δn, and follow $P_T(\Delta r)$. Below we provide all additional

information needed to calculate $\mathcal{L}(\{R,T\}|\mathbf{r},\mathbf{n})$. First we discuss how to determine the second distribution $P_{R1}(\Delta r_{i,j}|\Delta n_{i,j})$ and second the methodology for determining the two sets $\{R\ \&\ T\}$.

4.4.2.2 Determining $P_{R1}(\Delta r|\Delta n)$

To determine $P_{R1}(\Delta r|\Delta n)$ we utilize the pairwise distance distributions between localizations with a given frame $P_d(\Delta r|\Delta n)$ and the true pairwise distance distribution $P_T(\Delta r)$. Here $P_T(\Delta r)$ is known, determined using the pairwise distances between localizations that are separated by a frame greater than N (See Main Text).

Again, the desired distribution $P_{R1}(\Delta r|\Delta n)$ gives the probability of observing a distance between localizations for a given Δn if at least one of them is a repeat. $P_{R1}(\Delta r|\Delta n)$ is therefore made up of the distances between $\{R$ and $T\}$ and $\{R$ and $R\}$, where the curly brackets with the *and* indicate the pairwise distances between the localiztions within the sets. While $P_d(\Delta r|\Delta n)$ is made up of the distances between $\{R$ and $T\}$, $\{R$ and $R\}$, and $\{T$ and $T\}$ for a given Δn. Therefore, $P_{R1}(\Delta r|\Delta n)$ is equal to $P_d(\Delta r|\Delta n)$ with the contribution from the distances between true localizations removed, $\{T$ and $T\}$.

To properly eliminate the part of the distribution that is due to the distances between the true localizations, we quantify the makeup of $P_d(\Delta r|\Delta n)$ and then proportionally remove $P_T(\Delta r)$ from $P_d(\Delta r|\Delta n)$, resulting in $P_{R1}(\Delta r|\Delta n)$.

$P_d(\Delta r|\Delta n)$ is itself a combination of two distributions $P_T(\Delta r)$ & $P_{blink}(\Delta r)$, where the distances between different fluorophores follow $P_T(\Delta r)$ [Categories C1 and C2] and the distances between localizations from the same fluorophore follow $P_{blink}(\Delta r)$ [Category C3]. Here the probability distribution $P_{blink}(\Delta r)$ is the probability of observing a distance between a pair of localizations that are from the same fluorophore [Category C3] and is determined by the resolution of the SMLM experiment.

We can determine $P_{blink}(\Delta r)$ by comparing $P_T(\Delta r)$ to $P_d(\Delta r|\Delta n < N)$. The distribution $P_T(\Delta r)$ by definition lacks all distances between pairs of localizations that are from the same fluorophore and only contains distances between localizations from different fluorophores [Categories C1 and C2]. While $P_d(\Delta r|\Delta n < N)$ not only contains the distances between pairs of localizations from the same fluorophore [Category C3], but the distances between different fluorophores [Categories C1 and C2]. Note that within a SMLM experiment the resolution is very high, and therefore the distances between the localizations from the same fluorophore are very small, much less than 1000 nm. Therefore, the "shape" of the tails of the two distributions $P_T(\Delta r)$ and $P_d(\Delta r|\Delta n < N)$ match each other, as they both only contain the distances between different fluorophores (Data not shown). With this understanding in mind, the distribution $P_{blink}(\Delta r)$ can be obtained by subtracting $P_T(\Delta r)$ from $P_d(\Delta r|\Delta n < N)$ so that the probability of observing a distance greater than 1000 nm is approximately zero, and then normalizing so that the distribution

sums to one.

To determine the proportion of each distribution making up $P_d(\Delta r|\Delta n)$, $P_d(\Delta r|\Delta n)$ can be fit to the following equation:

$$X(\Delta n) = Fit[(1 - X) \times P_{blink}(\Delta r) + X \times P_T(\Delta r)], \tag{4.4}$$

where X is between 0 and 1.

The proportion of the distances that follow $P_T(\Delta r)$ come from the distances between $\{T$ and $T\}$ and $\{R$ and $T\}$. We must therefore take this into consideration when determining the proportion of $P_{R1}(\Delta r|\Delta n)$ that follows $P_T(\Delta r)$. To adjust the proportion of the distribution that follows $P_T(\Delta r)$ we calculate the ratio of the number of distances from $\{R$ and $T\}$ relative to the number of distances from $\{T$ and $T\}$ and $\{R$ and $T\}$.

This ratio can be determined by calculating the average number of repeats per fluorophore, num_b. num_b can be obtained without having to perform any additional experiments, using the approximate probability that a localization is a repeat (See Approximating the Probability a Localization is a repeat Section of this Supporting Material) and Alg. 1 to obtain a relatively accurate estimation as to the number of blinks per fluorophore. (Note: for this calculation $\kappa(density) = 0$ and $\kappa_2(frame) = 0$, discussed later.) Here we should note that num_b could also be determined by experiment, though these experiments can be difficult and are very sensitive to model specific errors.

The ratio of the number of distances from $\{R \text{ and } T\}$ relative to the number of distances from $\{R \text{ and } T\}$ and $\{T \text{ and } T\}$ is then the following (See Mathematical Justification Section of this Supporting Material):

$$\alpha = \frac{num_b + num_b + num_b * num_b}{1 + num_b + num_b + num_b * num_b} = \frac{\#\{R \text{ and } T\}}{\#\{R \text{ and } T\} + \#\{T \text{ and } T\}}. \tag{4.5}$$

where $\#\{R \text{ and } T\}$ indicates the number of distances between the localizations within the two sets. The distribution $P_{R1}(\Delta r|\Delta n)$ is then equal to the following equation:

$$P_{R1}(\Delta r|\Delta n) = Norm[P_T(\Delta r) \times X(\Delta n) \times \alpha + P_{blink}(\Delta r) \times [1 - X(\Delta n)]]. \tag{4.6}$$

Here $Norm$ indicates that the distribution within the brackets is normalized so that it sums to one. The distribution $P_{R1}(\Delta r|\Delta n)$ is a combination of the two distributions that are from the distances between localizations from different fluorophores ($P_T(\Delta r)$)) and the distances between the localizations from the same fluorophore ($P_{blink}(\Delta r)$). The first term ($P_T(\Delta r) \times X(\Delta n) \times \alpha$), first accounts for the proportion of the distribution $P_d(\Delta r|\Delta n)$ that results from the distances between localizations from different fluorophores and then scales this proportion further with α, so that the contribution from the distances between the pairs of true localizations are removed. $P_{R1}(\Delta r|\Delta n)$, for the 1 dark state no clustered simulation is shown in Fig. 4.8A. As expected, there

is a large probability for small distances and small frame differences due to the proportion of distances between the blinks of the same fluorophores being large. Then as the frame difference increases, the proportion of distances between the blinks of the same fluorophores decreases and the distribution converges upon the true pairwise distance distribution, Fig. 4.8A.

4.4.2.3 Determining the sets {R} and {T}

To assign a localization to either the {R} set (repeat) or the {T} set (True Localization) DDC uses the following:

$$\{R, T\} = Alg_1[\mathbf{r}, \mathbf{n}, \mathbf{M}, \kappa(density), \kappa_2(frame)]. \tag{4.7}$$

The sets {R} and {T} are determined within *Algorithm* 1, which uses the parameters and data within the brackets to assign each localization to one of the two sets. The actual experimental data are stored within the two terms **r** & **n**, where **r** contains the coordinates of each localization and **n** contains the frame. Here, **M** is a matrix that contains the information that is used to determine the probability that a localization is a repeat (See Approximating the Probability a Localization is a repeat Section) and $\kappa(density)$ & $\kappa_2(frame)$ are monotonic functions that are determined within the MCMC. The two functions $\kappa(density)$ & $\kappa_2(frame)$ allow DDC to adjust the probability calculation by taking into consideration the local density of the image and the frame of each localization. These are the two functions that vary during the MCMC

to maximize the likelihood, defining the two sets. We discuss the specifics of $\kappa(density)$ & $\kappa_2(frame)$ within the section Alg. 1, Linking Localizations into Trajectories.

4.4.3 Approximating the probability that a localization is a repeat

Depending upon the number of localizations within a SMLM image, the number of subsets of localizations can be extremely large. To speed up the phase space search and to minimize the likelihood of overfitting DDC calculates the approximate probability that each localization is a repeat (within the blinking trajectory) of a prior localization and only investigates the more likely subsets of localizations using the MCMC approach (Alg. 1, Linking Localizations into Trajectories). Below we discuss how the approximate probability that each localization is a repeat can be determined and then describe Algorithm 1, which defines which localizations are true localizations and which are blinks within DDC.

Here we define the matrix $\mathbf{M}$, which gives the probability that a localization is a repeat of a prior localization given a distance, Δr, and Δn between the localizations.

$$\mathbf{M}(\Delta r \in i, \Delta n) = \frac{P_d^i(\Delta r|\Delta n) - \omega \times P_T^i(\Delta r)}{P_d^i(\Delta r|\Delta n)}, \tag{4.8}$$

where $P_d^i(\Delta r|\Delta n)$ is the raw probability for the distance between two localizations to be in bin i, given that they are separated by Δn, $P_T^i(\Delta r)$ is the true

pairwise distance distribution and $\omega = \frac{\sum_{i>>\sigma} P_d^i(\Delta r|\Delta n)}{\sum_{i>>\sigma} P_T^i(\Delta r)}$, where σ is the localization precision of the microscope. Here ω is a scaling factor used to match the tails of the two distributions, as the distance distributions have a similar shape for $\Delta r >> \sigma$. Fig. 4.6 illustrates this calculation and the assumption about the tails of the distribution. **M**, for the 1 dark state no clustered simulation is shown in Fig. 4.8B, as expected, there are high probabilities with small Δr and small Δn, which get lower as Δr and Δn increase. **M** is the matrix that Alg. 1 uses to link localizations into trajectories.

4.4.4 Alg. 1, linking localizations into trajectories

Here we describe Alg. 1, which DDC uses to determine which localizations are linked into trajectories using the previously defined **M** and $\kappa(density)$ & $\kappa_2(frame)$. (See Approximating the Probability a Localization is a repeat) Note: one could easily modify the algorithm and have it take into consideration more information to determine which localizations belong to each set, but at a computational cost and risk of overfitting.

We wanted our methodology to be able to account for heterogeneous distributions of fluorophores within the same image and to incorporate the "time" dependence for the appearance of localizations. Therefore, one single cutoff probability or threshold was avoided. Instead we made the probability at which localizations are linked together into the same blinking trajectory related to the local density of the image before blinking correction and related to the frame of the localization.

Note: during the maximization of the likelihood for all of the systems within this work we could not simply eliminate localizations without taking into consideration the "probability of repeat", as this led to an extremely large phase space and did not converge within a reasonable amount of time.

Here the reasoning for incorporating the density is this: the more dense a region of an image is the more likely that a true localization could be considered a repeat by chance (based off of the probability calculation, see Alg. 1) and therefore the density of the image needs be taken into consideration. To incorporate the heterogeneity of the image DDC determines the local density of each localization before the blinking correction. To do this DDC calculates the number of raw localizations that are within 2σ (SMLM resolution) and have a frame difference greater than N, for each localization. DDC then defines a monotonically increasing function that is a function of the density, $\kappa(density)$ [Initially $\kappa(density) = 0$]. The flexibility of this function allows DDC to handle heterogeneous distributions of fluorophores by taking into consideration the local density of the image for the probability calculation (See Alg. 1).

Note: the shape of this function is determined during the MCMC approach and is discussed within Alg. 2.

The reasoning to include the frame information within the probability calculation is: because more localizations appear at the beginning of the acquisition

of an image when compared to the end of the acquisition, localizations would be more likely to be considered repeats at the beginning of the acquisition than at the end by random chance. (Because fluorophores photo-bleach during the acquisition of a SMLM image.) The time dependence is utilized in a similar manner as the density, where a monotonically decreasing function of the frame of each localization is incorporated into the probability calculation, $\kappa_2(frame)$, see Alg 1.

Note: the shape of this function is also determined during the MCMC approach and is discussed within Alg. 2.

To link localizations into trajectories DDC utilizes Alg. 1. This simple algorithm goes through all localizations and links them into trajectories, starting with the localizations that are most similar in frame. To decide whether or not to link two localizations into the same trajectory [or two trajectories into one] the algorithm used the mean of the "probabilities of blink" of the localizations being considered. DDC calculates the probability of being a blink with the matrix $\mathbf{M}$, and then divides the mean probability by $1 + \kappa(density(ii)) + \kappa_2(frame(ii))$. This takes into account the local density and frame of the localization ii. If the probability of the localization [or localizations] is larger than .5, then the localizations are combined into the same trajectory. For each trajectory all localizations but the localization with the smallest frame in each trajectory are then considered blinks.

Note: we should mention that the order in which the localizations in Alg. 1 are arranged does have a small influence on the trajectories that are formed, especially if the activation rate is high. Therefore, DDC also varies the order of the localizations during the MCMC approach to obtain different subsets of true localizations (See Alg. 2 of this Supporting Material for further details.)

Note: we found that not including an algorithm of similar structure to Alg. 1 (takes into account the physical process of fluorophore blinking) either resulted in an extremely slow convergence or got stuck in minimums that deviated from the true image. Therefore, including the information within $\mathbf{M}$ is critical for DDC to converge upon the true image. We should also state that we did not perform an extensive search for alternatives and we do realize that improvements to Alg. 1 could be an area of improvement for DDC in future research.

4.4.5 Alg. 2, MCMC approach to maximize the likelihood

Here we describe Alg. 2, which DDC uses to maximize the Likelihood and obtain the "correct" subset of true localizations.

Algorithm 2 is a simple Markov Chain Monte Carlo (MCMC) approach that utilizes Alg. 1 in the process. The MCMC approach perturbs three parameters, $\kappa(density)$, $\kappa_2(frame)$ and the order of the localizations to determine the "correct" subset of true localizations. For each step, one of the three previous parameters are modified by a small amount and the likelihood is calculated for

the particular subset of true localizations determined by Alg. 1, given those parameters. Alg. 2 then keeps the new parameter and resets the best likelihood if the likelihood is greater than the previous best likelihood or accepts the new parameters if the difference of the likelihood with the old likelihood is greater than a uniform random number. An example of a phase space search is shown in Fig. 4.9, where the maximization of the likelihood results in the results shown in red.

We found that including the MCMC approach to maximize the log of the likelihood led to significant improvements in the correct number of fluorophores calculated for all systems. Furthermore, for the more heterogeneous distributions of localizations, the Small clusters simulation systems, the MCMC approach led to dramatic improvements in the image error, data not shown. Therefore, the MCMC approach is vital to the successful supplication of DDC even though it is the most computationally expensive step of the methodology.

4.4.6 Evaluating the three most common threshold methodologies and the absolute best image error from thresholding

Here we investigate the three most common threshold methodologies and compare their results with DDC. We also compare DDC to the absolute best Image Error thresholding can produce. We discuss the results from each of the comparisons here and whenever we reference the 2 dark state systems we are referring to Fig. 2 in the main text and whenever we mention the 1 dark

state system we are referencing the results shown in Fig. 4.14.

4.4.6.1 Equations for evaluating the different methods

The image error of each methodology was calculated with the following equation:

$$ImageError = \frac{\sum_{i,j}[Norm(CorrectedImage(i,j)) - Norm(RealImage(i,j))]^2}{\sum_{i,j}[Norm(UncorrectedImage(i,j)) - Norm(RealImage(i,j))]^2}, \tag{4.9}$$

where $i\&j$ go over all pixels within the images, Norm() indicates that the image is normalized so that the maximum intensity is 1 and the lowest intensity is 0, CorrectedImage is the image that results from a blinking corrected methodology, RealImage is the image that results if an image is generated only using the true localizations and UncorrectedImage is the image with no blinking corrected methodology.

The counting error of each methodology was calculated with the following equation:

$$CountingError = \frac{|Methods\#ofloc - Real\#ofloc|}{Real\#ofloc} \times 100, \tag{4.10}$$

where $Methods\#ofloc$ is the number of true localizations determined by the methodology and $Real\#ofloc$ is the actual number of true localizations.

4.4.6.2 2011, Semi-empirical equation to obtain photo-kinetics (T1)

Perhaps the most famous and most widely used methodology to extract the photo-kinetics and correct for blinking is by utilizing a semi-empirical formula

developed in 2011 [181]. The parameters from the fit to the semi-empirical formula are also often used with the suggested optimal thresholds from Coltharp et al. [178] with a time threshold equal to 2 times the average dark time of the fluorophore.

For this methodology the distance threshold is often set to 1 pixel (100nm) and the time threshold, t_d (dark time) is varied and the number of localizations at each t_d is quantified. Once the longer t_d is determine the time threshold is often set to approximately 2 times the dark time.

To evaluate the effectiveness of this methodology we applied the methodology to the 1 dark state simulation data for the three different distributions of fluorophores, Fig 4.12. The semi-empirical formula fit well, but the error in the number of fluorophores and the average dark time was very significant, indicating that the methodology is flawed for systems with more than 1 blink per fluorophore. (The percent error for the extracted parameters is shown in the titles of each subplot.) This previously unknown degree of error is likely due to the small number of simulation systems to which the methodology was applied during the development of the methodology. Though, here we feel that we should state that this previous work was vital for informing the field just how important blinking correction can be.

Considering the large amount of error in the extracted parameters, Fig. 4.12, we choose to assume that the methodology had perfect knowledge of the char-

acteristic times for the dark states for each simulation system. When comparing the error with the time threshold set to 2 times the known dark time the error in the calculated number of fluorophores improved significantly, Fig. 4.14. For the two dark state simulation data we set the time threshold equal to 2 times the longer characteristic dark time. The results of applying these thresholds are shown in Fig. 2 in the main text.

When compared to DDC across all molecular distributions and fluorophores DDC outperformed this methodology across every metric. Considering this is the only other methodology that does not require additional experiments to quantify the photo-kinetics of the fluorophore, the experiments here suggest that DDC should be utilized in every situation instead of this methodology.

4.4.6.3 2013, Stringent thresholds to eliminate possibility of over-counting (T2)

For the thresholding methodology of Puchner et al. [180], they first characterized the photo-kinetics of the fluorophore and then set an extremely stringent time threshold, so that 99% of blinking dark times would be linked together and a distance threshold equal to 4 times the resolution of the experiment. This methodology was mainly developed to eliminate the possibility of blinking leading to the appearance of clusters, but due to the extreme thresholds this method will deplete the intensity of true clusters.

The results of comparing this thresholding methodology to the 1 dark state simulation systems is shown in Fig. 4.14. For the Image Error in each of the

3 systems DDC was significantly better than this thresholding methodology. The improvement was especially noticeable for the dense 1 dark state system, as the stringent thresholds are expected to be detrimental to dense clusters. Suggesting that DDC is better at obtaining the true underlying distribution of fluorophores.

Interestingly, this methodology performed especially well for the number of fluorophores in the random and Small clusters 1 dark state systems, but failed for the dense system with a percent error around 15%. When compared to DDC for the number of fluorophores, DDC consistently had a percent error less than 5%. Suggesting that DDC is also a more reliable method under this metric for this fluorophore.

The results for comparing this thresholding methodology with DDC for the 2 dark state simulation systems is shown in Fig. 2 in the main text. Across the board DDC was vastly better than this thresholding methodology for both the Image Error and the error in the number of fluorophores. Suggesting that when the photo-kinetics of the fluorophore are more complicated than a simple 1 dark state DDC is especially beneficial when compared to this methodology. Furthermore, this thresholding methodology requires the characterization of the fluorophore, which wastes valuable time and can be experimentally difficult at times.

4.4.6.4 2012, Determining thresholds by knowing the number of fluorophores (T3)

In the methodology developed by Coltharp et al. [178] they characterized the fluorophores to determine the number of blinks per fluorophore to determine the time threshold and the distance threshold. To determine the number of blinks per fluorophore Coltharp et al. utilized a low activation (UV) laser and slowly activated the fluorophores so that individual time traces could be easily extracted. In the last section of the results of the main text we show that this methodology is likely flawed and varying activation intensities change the photo-physics of the fluorophores potentially leading to errors in the number of blinks per fluorophore, Fig. 4. Though, further experiments would be needed for that particular fluorophore. Also, even if the time traces are properly extracted from fluorophores with the same photo-physics the fits to the dark time intervals are error prone and model dependent [178].

Assuming perfect knowledge as to the number of blinks per fluorophore for this methodology, we scanned the number of localizations obtained for each time threshold and distance threshold. The ideal thresholds were determined using the thresholds for the minimal error in the number of localizations at the intersection of the time and distance thresholds. Examples of this phase space search for six different systems investigated in this work are shown in the first column of Fig. 4.13, with the corresponding Image Error for each set of thresholds shown in the second column. (Note: the error is log scale for the first column so one can clearly see why the exact thresholds were chosen.) The

thresholds determined by this methodology are shown in the following table:

System	Time Threshold (n)	Distance Threshold (nm)
Random 1 dark	25	130
Small clusters 1 dark	20	130
Dense 1 dark	20	100
Random 2 dark	35	130
Small clusters 2 dark	35	130
Dense 2 dark	30	100
Filament	35	80

Note: Logically, the optimal thresholds for this methodology became less intense the more dense the molecular distributions became.

The results of applying this methodology are shown in Fig. 4.14 for the 1 dark state systems and Fig. 2 for the 2 dark state systems. Considering with this methodology we assumed perfect knowledge for the number of blinks per fluorophore it was of little surprise that the error in the calculated number of fluorophores was actually lower than DDC for the 1 dark state systems. The error in the number of fluorophores was less than 6% for both methods for all systems for the 1 dark state fluorophore. Even though the error in the number of fluorophores for both methodologies was comparable, the DDC Image Error was lower for each 1 dark state system when compared to this thresholding methodology. Suggesting, that DDC captures a more reliable representation of the true localizations, while resulting in a comparable error in the number of fluorophores for the simple 1 dark state fluorophore.

This was also the case for the 2 dark state simulation systems except for the dense distribution system. For the dense distribution system the error in the number of fluorophores was significantly worse for DDC, about 12%, while the thresholding methodology performed well with this metric. (We should note again that this is assuming perfect knowledge as to the number of blinks per fluorophore, so it is expected that the error in the number of fluorophores will always be low with this methodology.) Even though DDC performed worse for the dense 2 dark state system for the number of fluorophores, for the Image Error DDC greatly surpassed this thresholding methodology for all three distributions of fluorophores. The most significant improvement was for the dense system, where this thresholding methodology performed much worse than even an uncorrected SMLM image. Suggesting that DDC is vastly superior than this thresholding methodology for a more complicated 2 dark state fluorophore and great care should be taken when utilizing this methodology when actual clustering exists.

4.4.6.5 The absolute best thresholds for the image error (T4)

Considering DDC was able to surpass all of the traditional thresholding methodologies with regards to the Image Error, we wanted to see if any thresholds could surpass DDC. To do this we scanned the time threshold and distance threshold for each system and picked the thresholds that resulted in the mean minimum Image Error for each of the seven systems. The thresholds picked by this methodology are shown in the following table:

System	Time Threshold (n)	Distance Threshold (nm)
Random 1 dark	17	160
Small clusters 1 dark	13	170
Dense 1 dark	5	190
Random 2 dark	39	140
Small clusters 2 dark	28	150
Dense 2 dark	3	210
Filament	43	80
Continuous Filaments	10	80

The results of comparing the absolute best threshold methodology with DDC is shown in Fig. 4.14 for the 1 dark state system. As expected this thresholding methodology performed best for the metric of Image Error when compared to the other thresholding methodologies. Interestingly, DDC was still able to outperform the thresholding methodology in terms of the Image Error and in terms of the number of fluorophores.

The results of comparing this thresholding methodology with DDC for the 2 dark state system is shown in Fig. 2. Interestingly, for this fluorophore the Image Error and the error in the number of fluorophores for the Random and the Small clusters systems was similar between the two methods. The major difference was for the dense system where the error in the number of fluorophores was around 80% for the thresholding system, while DDC maintained an error of about 12%. Suggesting that the Image Error for the 2 dark state systems was similar between the two methods, but DDC was able to surpass

this thresholding methodology in terms of determining the proper number of fluorophores.

These results suggest that even with the absolute best thresholds DDC is still a more reliable approach in regards to the two metrics investigated within this work.

4.4.7 Methodology of Sphan et al.

The implementation of Sphan et al. was done by randomly selecting subsets of localizations (with replacement) and then using the threshold of 2.5 (just as in [190]) as the definition of a cluster — to create the cluster masks. The normalized average density within the clusters (P/P_o) vs. the relative area of the image the clusters covered (η) was plotted for all subsets of localizations to determine if clustering was significant for the system of interest. For this methodology, clustering is deemed significant if P/P_o rises above 1 and stays above 1.

We tested this method on three different simulation systems (Random, Small Clusters, Dense Clusters) with the two-state fluorophore and show these results in Fig. 4.20A. We observed that the randomly distributed fluorophores maintained a P/P_o equal to 1 while the Dense cluster system rose significantly well above 1, demonstrating that the methodology could adequately recognize that there were clusters in the Dense cluster system and that there were not clusters in the Random system. As expected an intermediate value for the

Small cluster system was also observed.

Next, to investigate the clustering of AKAP79/150 with an orthogonal method to DDC, we applied the methodology of Spahn et al. on the superresolution data of each of the two orthologs. The results of this analysis are shown in Fig. 4.20B, where P/P_o for both rose slightly above a $P/P_o = 1$. These results support the previous findings that the two are significantly clustered, supporting the analysis as quantified by DDC. Though, we should note that P/P_o did not reach high values (like that for the Dense cluster system), suggesting that just as with DDC, the clustering of the two orthologs are not "extreme."

4.4.8 Specifics for simulations

First, six different sets of data were simulated, 3 different underlying structures and 2 different fluorophores. The two fluorophores followed the two models in Fig.4.3. In these simulations the fluorophores only registered as a localization if it was in the active state. For the different simulations the first condition contained no clusters [Random] and all fluorophores were randomly distributed within a 1000nm by 1000nm square and allowed to blink according to the kinetic models in Fig.4.3. The second [Small clusters] and third [Dense] conditions had 3 clusters each with 10% of the fluorophores distributed into the clusters for the Small clusters system and 50% for the Dense system. For each of the simulations with clusters each cluster's central location was randomly defined and the localizations within each cluster followed

a normal distribution around that center with a $\sigma = 40$. For each of the six systems 24 different images were generated and analyzed for each methodology.

Second, for the simulations involving filaments, we randomly distributed 50% of the true localizations along 5 lines and randomly deposited the rest randomly. We simulated 24 images, with 1000 true localizations each, with approximately 4000 localizations total, following the photo-kinetic model in Fig.4.3A. These simulations produced filaments that were clearly visible, but not homogeneous along the filaments.

Third, to produce "regular" continuous overlapping filaments we simulated filaments with no varying label density and with a localization error of 20 nm. This was done by placing a fluorophore every 5 nm along a filament. These simulations also followed Fig.4.3A and resulted in images like that in Fig. 4.16.

4.4.9 Methods for experiments that were used to calculate $Z(\Delta n)$

4.4.9.1 Strains

The strains with chromosomal fluorescent protein fusion tags were constructed using λ-RED-mediated homologous recombination [198]. Some results used in this paper came from strains that also harbor a single chromosomal DNA site marker (tetO6), the DNA markers are positioned in various positions on the chromosome, and a portion of the results are not relevant and thus not discussed in this publication. The details for the construction of these bacterial strains are described in detail in a previous publication [198].

4.4.9.2 Cell growth

For live cell imaging, single colonies were picked from LB plates and cultured overnight in EZ Rich Defined Media (EZRDM, Teknova) with 0.4% glucose, at room temperature (RT) with shaking. The next morning, cells were reinoculated into fresh EZRDM with 0.4% glucose and grown at RT until they have reached mid-log phase (O.D.600 0.3-0.4). For simultaneous visualization of DNA site markers (results are not reported here), cells were harvested and resuspended in fresh EZRDM supplemented with 0.3% L-arabinose and 0.4% glycerol and allowed to grow for two additional hours, these cells were harvested via centrifugation and imaged immediately.

For fixed cell experiments, cells were grown accordingly and fixed in 3.7% (v/v) paraformaldehyde (16% Paraformaldehyde, EM Grade, EMS) for 15 min at RT, washed with 1X PBS and imaged immediately.

4.4.9.3 Nascent rRNA labeling (smFISH)

We performed smFISH using a previously published protocol ([199], [200]). Briefly, cells were grown in EZRDM glucose as previously described; 5 ml of mid-log phase cells were fixed with 3.7% (v/v) paraformaldehyde (16% Paraformaldehyde, EM Grade, EMS), placed for 30 min on ice. Next, cells were harvested via centrifugation, and subsequently washed two times in 1X PBS. Cells were then permeabilized by resuspending in a mixture of 300 l of H2O and 700 μl of 100% ethanol and incubating with rotation at RT for 30

min. Cells were stored at 4 °C until next day. Wash buffer was freshly prepared with 40% formamide and 2x SSC and put on ice. Cells were spun-down in a bench-top centrifuge at 10000 rpm for 3 min and the cell pellet was resuspended in 1 ml of wash buffer. The sample was placed on a nutator to mix for 5 min at RT. Hybridization solution was prepared with 40% formamide and 2x SSC, subsequently, dye-labeled oligo probes were added to hybridization solution to a final concentration of 1 μ M. Cell were spun-down again and 50 μl of hybridization solution with probe was added to the pellet. The hybridization sample was mixed well and placed overnight in a 30 °C incubator. Next day, 10 μl of hybridization sample was washed with 200 μl of fresh wash buffer and incubated at 30 °C for 30 min, this was repeated one more time. The washed sample was imaged immediately: without STORM imaging buffer for ensemble fluorescence, with STORM buffer to induce dye blinking for superresolution imaging. glucose oxidase + thiol STORM buffer was used to image samples with only dye labeling (50 mM Tris (pH 8.0), 10 mM NaCl, 0.5 mg ml-1 glucose oxidase (Sigma-Aldrich), 40 g ml-1 catalase (Roche), 10% (w/v) glucose and 10 mM MEA (Fluka))([201]). Thiol only STORM buffer (10 mM MEA, 50 mM Tris (pH 8.0), 10 mM NaCl) was used to image samples with both endogenously expressed fluorescent proteins and dye labeling. This was to preserve the fluorescent signal from fluorescent proteins, since the presence of glucose oxidase in the STORM buffer tended to quench the fluorescent protein signal. Pre-rRNA transcripts were detected with a single probe L1, conjugated at the 5' with either Alexa Fluor 488 (NHS ester) or Alexa Fluor 647 (NHS ester) (IDT) ([202]). Upon receiving the commercial oligos, a

working stock (50 M) was made and aliquoted for storage at -20 °C.

4.4.9.4 Cell imaging and SMLM analysis

A 3% gel pad made with low-melting agarose (SeaPlaque, Lonza) in EZRDM was prepared. Live cells of an optimal imaging density were deposited onto the gel pad and immobilized with a coverslip for imaging as previously described ([199]). An Olympus IX-81 inverted microscope with a 100X oil objective (UPlanApo, N = 1.4x) was used, with 1.6x additional amplification. Images were captured with an Ixon DU-895 (Andor) EM-CCD with a 13 μm pixel size using MetaMorph (Molecular Devices). Illuminations (405 nm, 488 nm, 561 nm, 647 nm) were provided by solid-state lasers Coherent OBIS-405, Coherent OBIS-488, Coherent Sapphire-561, and Coherent OBIS-647 respectively. Fluorescence was split using a multi dichroic filter (ZT 405/488/561/647rpc, Chroma), and the far-red, red and green channels were further selected using HQ705/55, HQ600/50 and ET525/50 bandpass filters (Chroma). Gold fiducial beads (50 nm, Microspheres-Nanospheres, Mahopac, NY) were used to correct for any sample drift during imaging. All superresolution images were acquired with a 10 ms exposure time with 3000-9000 frames. Activation of fluorescent proteins was done simultaneously to fluorophore excitation, and activation laser (405) was kept at a constant power throughout the imaging session. For two-color imaging, the simultaneous, multi-color acquisition was achieved using Optosplit II or Optosplit III (Cairn Research), colored channels were overlaid using calibration images from TetraSpeck beads (Life Technologies, T-7279), as previously described ([203]). Initial fitting of raw imaging data was performed via thunderSTORM plugin ([204]). Later analysis of lo-

calizations with DDC was processed using custom Matlab scripts, which will be made available upon request.

4.4.10 Methods used for sister chromatid experiments

4.4.11 Methods used for dynein experiments

4.4.11.1 CELL LINE

Stably transfected HeLa IC74-mfGFP cells (The dynein intermediate chain is GFP labeled, from Takashi Murayama lab, Juntendo University School of Medicine, Tokyo, Japan) were plated on a 8-well Lab-tek 1 coverglass chamber (Nunc). Cells were cultured under standard conditions (DMEM, high glucose, pyruvate, 10% FBS and 2 mM glutamine).

4.4.11.2 IMMUNOSTAINING

Cells were fixed with PFA (4% in PBS) at RT for 20' and incubated with blocking buffer (3% (wt/vol) BSA (Sigma) in PBS and 0.2% Tryton X-100 (Thermo Fisher Scientific) for 1 hr. Dynein intermediate chain-GFP was immunostained with primary antibody (chicken polyclonal anti GFP, Abcam 13970) diluted 1:500 in blocking buffer for 45 minutes at RT. Cells were rinsed 3 times in blocking buffer and incubated for 45 minutes in secondary antibodies donkey-anti chicken labeled with photoactivatable dye pairs for STORM (Alexa Fluor 405-Alexa Fluor 647).

4.4.11.3 IMAGING

Imaging was done using Nanoimager-S microscope (Oxford Nanoimaging) with the following specifications: 405, 488, 561, and 640 nm lasers, and 665–705 nm band-pass filters, 100× 1.4 NA objective (Olympus), and a Hamamatsu Flash 4 V3 sCMOS camera. Localization microscopy images were acquired with 16-ms exposure and 50,000 frames. 405-nm activation was kept constant and then processed using the NimOS localization software (Oxford Nanoimaging).

4.4.12 Methods used for AKAP150

For fixed-cell stochastic optical reconstruction microscopy (STORM) imaging, cells were fixed with 4% paraformaldehyde (PFA) for 20 min and then washed with 100 mM glycine in Hanks balanced salt solution (HBSS) to quench the free PFA. Cells were permeabilized and blocked in a permeabilization solution with 0.1% Triton X-100, 0.2% bovine serum albumin, 5% goat serum, and 0.01% sodium azide in HBSS. The cells were then incubated overnight at 4°C with an anti- AKAP150 (Millipore Sigma 07-210, Cat. # 07-210 EMD Millipore) antibody at a 1:500 dilution, followed by 1 to 2 hours with goat anti-rabbit Alexa 647?conjugated antibodies at 1:1000 dilution. The cells were then post-fixed again in 4% PFA, quenched with 100 mM glycine in HBSS, and washed with HBSS to prepare for imaging. Immediately before imaging, the medium was changed to STORM-compatible buffer [50 mM tris-HCl (pH 8.0), 10 mM NaCl, and 10% glucose) with glucose oxidase (560 mg/ml), catalase (170 mg/ml), and mercapto-ethylamide (7.7 mg/ml). STORM images were obtained using a Nikon Ti total internal reflection fluorescence (TIRF) micro-

scope with N-STORM, an Andor IXON3 Ultra DU897 EMCCD, and a 100x oil immersion TIRF objective. Photoactivation was driven by a Coherent 405-nm laser, while excitation was driven with a Coherent 647-nm laser. Puncta localization was performed using both Nikon Elements analysis software.

4.4.13 Methods used for characterizing blinking

4.4.13.1 Sample preparation:

Plac::mEos3.2 plasmid (pXY329) was constructed based on pJL005 (Plac::FtsZwt-mEos3.2) [205] using In-fusion cloning (Takara) to remove the ftsz gene. MG1655 cells were transformed with pXY329 and grow up in M9+ media. The cells are harvested at log-phase and fixed by 3.8% para-formaldehyde in 1X PBS buffer. The fixed cells were washed by 1X PBS for 3 times and saved in 4°C no longer than one week.

Streptavidin conjugated with $AlexaFluor^{TM}647$ (SA-AF647) was purchased from Thermo Fisher Scientific. The SA-AF647 working solution was made freshly every time by diluting original stock (36μM) to 10 pM in 1X PBS with 0.5% Tween20.

4.4.13.2 Imaging

PALM: Fixed MG1655-Plac::mEos3.2 cells were sandwiched between a 3% PBS agar-pad and a coverglass as previously described [44]. PALM imaging was preformed as previous study [205] on an Olympus IX71 inverted micro-

scope with a 100X, 1.49 NA oil-immersion objective. The 561nm excitation laser power was tuned to 1500 W/cm2 while the 405nm laser power varied from 0 to 3.5 W/cm2. For the 0 W/cm2 condition, a short pulse (1 second) of 3.5 W/cm2 405nm laser was applied to activate some mEos3.2 molecules to red fluorescent state. At each 405-power condition, 6 movies of 3000 frame images with 10ms exposure time were collected continuously. Three repeats of all the 405-conditions were performed to get the average blinking behavior.

dSTORM: 10pM SA-AF647 was flown into a preassembled chamber with biotin-PEG coated coverglasses from X for 5min and washed three times with 1X PBS. The STORM buffer was made freshly using the recipe described in [206] and injected to the chamber to replace the PBS buffer before imaging. All STORM images were taken after 60 min since the oxygen level in the buffer was shown to be stable after 1 hour. dSTORM imaging was performed on an Olympus IX81 inverted microscope with a 100X, 1.45 NA oil-immersion objective. The 647nm excitation laser power was tuned to 1800 W/cm2 while the 405nm laser power varied from 0 to 13.9 W/cm2. At each 405-condition, 2-3 5000-frame movies at different regions on the coverglass were taken with a 30ms exposure time. Two repeats of all the 405-conditions were performed.

4.4.13.3 Data processing

The single fluorophore spots in both PALM and dSTORM movies were localized by an ImageJ [207] plugin ThunderSTORM [153]. All the spots with irregular properties (abnormal sigma, too high or low intensity, or multiple

spots within 500 nm range) were removed. A customized Matlab code was used to link the same spots within 3-4 folds of localization limitation (100nm) throughout the whole movie using a nearest neighbor algorithm. The continuous frames with localization from the same linked fluorophore were counted as on frames. Other frames before the last on-frame were counted as off frames. Blinking number was calculated as the sum of on frame number.

4.4.14 Algorithms

Algorithm 1

procedure DETERMINE WHICH LOCALIZATIONS ARE BLINKS

$\mathbf{M}(\Delta r, \Delta n) \leftarrow$ Probability that a localization is a repeat of the preceding localization given the Distance and Frame between the preceding localization

$traj(i) \leftarrow$ is the trajectory that localization i is assigned (before the for loops each localization is assigned it's own personal trajectory)

$\Delta\mathbf{r}_{traj(i),traj(ii)}$ and $\Delta\mathbf{n}_{traj(i),traj(ii)} \leftarrow$ arrays containing the pairwise distances and frame differences between all localizations in the two trajectories containing localization(i) and localization(ii)

$\Gamma = length(\Delta\mathbf{r}_{traj(i),traj(ii)})$

$\kappa(density(i)) \leftarrow$ a monotonically increasing function that is dependent upon the local density of localization(i) without blinking correction (Supporting Material).

$\kappa_2(frame(i)) \leftarrow$ a monotonically decreasing function that is dependent upon the frame of localization(i) (Supporting Material).

{T}=1:length(Localizations) $\leftarrow$ the indices that are the true localizations

{R}=empty array $\leftarrow$ the indices of the localizations that are repeats

for Δn=1:max(frame) **do**

  for i=1:length(Localizations) **do**

    for ii={T} **do**

      if frame(ii)-frame(i)=Δn **then**

        if $\frac{[\sum_j^{\Gamma} M(\Delta\mathbf{r}_{traj(i),traj(ii)}(j),\Delta\mathbf{n}_{traj(i),traj(ii)}(j)]/\Gamma}{1+\kappa(density(ii)+\kappa_2(frame(ii))} > .5$ **then**

          Combine all the Localizations within the two trajectories into a single trajectory

          Eliminate Localization(ii) from {T} as it is now considered a repeat

          Include Localization(ii) in {R} as it is now considered a repeat

Algorithm 2

procedure MARKOV CHAIN MONTE CARLO TO MAXIMIZE LIKELIHOOD
    Max Lik=$-\infty$
    Count=1
    Number of Steps=1000
    while Count<Number of Steps **do**
        $\kappa(density(:)) = \kappa^{Stored}(density(:))$
        $\kappa_2(frame(:)) = \kappa_2^{Stored}(frame(:))$
        $C = rand \leftarrow$ a random uniform number
        if $C < 1/3$ **then**
            Adjust the function $\kappa(density(:))$ by a small amount
            Ensure that $\kappa(density(:))$ is still a monotonically increasing function of density
            Ensure that the mean of $\kappa(density(:))$ over all density values from all localizations equals zero
        else if $C < 2/3$ **then**
            Adjust the function $\kappa_2(frame(:))$ by a small amount
            Ensure that $\kappa_2(frame(:))$ is still a monotonically decreasing function of the frame
            Ensure that the mean of $\kappa_2(frame(:))$ over all localizations equals zero
        else
            Perturb the order of localizations that have the same frame ▷ This will change which localizations are linked together into the same trajectory
        $\{R, T\} \leftarrow$ Perform Alg. 1 with new $\kappa(density(:))$, new $\kappa_2(frame(:))$, and in new defined order
        Lik $\leftarrow$ Calculate log likelihood with new Corrected Localizations
        if Lik>Max Lik or $log(rand) < |MaxLik - Lik|$ **then**
            Store new parameters
            Max Lik=Lik
        else
            Go back to old parameters
        Count=Count+1

Chapter 5

A Biophysical Model of Supercoiling Dependent Transcription Predicts a Structural Aspect to Gene Regulation

[1]

5.1 Background

The dynamics of gene expression in single cells has been studied extensively in the last 15 years, yielding new insights into the processes of transcription and translation [208, 209, 210, 211, 212, 213]. Populations of cells are now known to exhibit a large degree of heterogeneity in both mRNA and protein expression levels [214]. The probabilistic nature of molecular reactions gives rise to the intrinsic component of this variation, while the differences between cells,

[1] Bohrer CH, Roberts E. A biophysical model of supercoiling dependent transcription predicts a structural aspect to gene regulation. BMC biophysics. 2016 Dec;9(1):2

such as the levels of RNAP, ribosomes, etc, produce the extrinsic component [215, 216]. Both types of noise contribute to the wide distributions of mRNA and proteins in a population [217, 218, 219, 220, 221]. But only extrinsic fluctuations are typically considered capable of introducing correlations into the expression levels of different genes within a single cell. The fluctuations and correlations can in principle be used to study the details of the underlying molecular processes [222].

One repeated theme when studying single cell gene expression is the occurrence of bursts during the production of mRNA and/or proteins. In particular, the production of mRNA has been shown to deviate from a simple birth death process, instead occurring in transcriptional bursts [223, 224]. Transcriptional bursting gives rise to distributions with a Fano factor factor greater than one [223, 225, 221, 226]. Furthermore, it has been shown through single-molecule mRNA studies that the transcriptional bursting behavior is dependent on the expression levels and promoter architectures of the genes [225, 221]. The origin of these transcription bursts is still a subject of active inquiry and debate [221].

Recently, Chong et al. provided evidence supporting a possible mechanism for transcriptional bursting in *E. coli* [82]. As RNA polymerase (RNAP) translocates along the DNA producing mRNA, positive supercoiling is generated downstream and negative supercoiling upstream of the enzyme complex [227]. In the absence of other factors, dissociation of RNAP would enable the positive and negative supercoils to resolve each other, leaving a zero net change in the supercoiling state. In *E. coli* there are two major factors when it comes

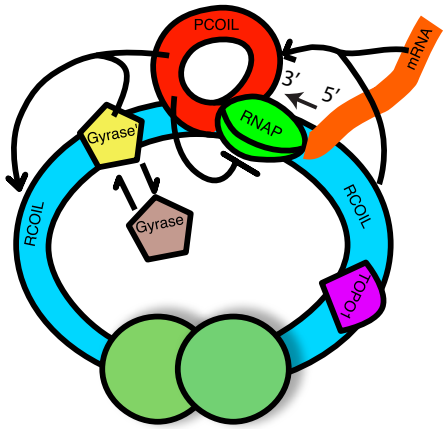

Figure 5.1: Positive Supercoiling (Pcoil) is produced when mRNA is transcribed. Pcoil inhibits the production of mRNA by reducing the initiation rate. In order to relieve Pcoil gyrase must bind (Gyrase'), which converts Pcoil into the "regular" state (Rcoil).

to relieving supercoiling generated by transcription, topoisomerase I (Topo I) and gyrase. Topo I relieves negative supercoiling while gyrase relieves positive supercoiling. In *E. coli* topo I has a higher activity than gyrase, as negative supercoiling can be very detrimental to the organism. This imbalance causes positive supercoiling to accumulate until gyrase binds and relieves the positive supercoiling [82]. Chong et al. provided evidence that the build-up of positive supercoiling was the underlying mechanism for transcriptional bursting, where transcriptional bursts happen when positive supercoiling that inhibits the initiation rate of the gene is relieved, see Fig. 5.1 .

Many models of gene expression have been proposed and studied, *e.g.* see [228, 229, 230, 231]. However, none of these models account for the generation of positive supercoiling from transcription events. If this is part of the mechanism by which transcriptional bursting takes place, incorporating the

accumulation of positive supercoiling in gene expression is vital in order to correctly describe the fluctuations and correlations of the system.

Here, we first develop a biophysical model to quantify the effect of supercoiling density on the transcription initiation rate. Then, using a simplified version of this model, we create a kinetic model of gene expression that directly incorporates the accumulation of supercoiling. We found that only when the supercoiling was accounted for in the model were we able to qualitatively reproduce the diverse distributions of mRNA seen in experimental studies with *E. coli* [82, 214]. We then investigated the effect of having multiple genes in the same supercoiling domain and find there to be a correlation in the expression of these genes. Having multiple genes in the same supercoiling domain also results in each gene's expression influencing the expression of other genes in the same domain. These results not only provide vital insight as to how different genes are expressed and regulated in bacteria, but also provides new directions for experimentally testing for the effects of domain coupled transcriptional bursting.

5.2 Biophysical model for RNAP initiation with supercoiling

In order to produce mRNA, RNAP must bind and melt the DNA strands to allow an RNA-DNA hybrid to form before proceeding to elongation [225, 232]. This necessitates to maintaining the stability of this open complex long enough to form the DNA-RNA hybrid so RNAP can form a stable elongation complex

[232, 233]. Recently, Chong et. al. utilized an *in vitro* assay to demonstrate the effect of positive supercoiling on the rate of initiation for RNAP. Their experiment monitored the production of mRNA from 160 individual molecules using an RNA-specific fluorescent dye. In the absence of gyrase, positive supercoiling accumulated, resulting in a decrease in the initiation rate of transcription. In their experiments, this manifested as a decrease in the fluorescence intensity with time as mRNA transcripts were being produced less frequently (Fig. 5.2 A). The cumulative sum of this data equals the total number of transcription events that had occurred by a given time (Fig. 5.2 B, red line). Using the reported intensity of a single mRNA transcript, 13.5×10^3 [82], the average time it takes every template to produce an mRNA molecule can be calculated; this is termed an "average transcription event" (Fig. 5.2 B, green triangles). The time between successive transcription events shows the decline in initiation rate due to positive supercoil accumulation (Fig. 5.2 C). The decrease in the initiation rate with each average transcription event can be seen in Fig. 5.2 D and roughly decreases linearly.

In order to study why the initiation rate decreases in this manner we use the following kinetic model:

$$\mathrm{RNAP} + \mathrm{DNA} \underset{k_{\mathrm{off}}}{\overset{k_{\mathrm{on}}}{\rightleftharpoons}} \mathrm{RNAP} : \mathrm{DNA_c} \overset{k_{\mathrm{cat}}}{\rightarrow} \mathrm{RNAP} : \mathrm{DNA_o} \overset{k_{\mathrm{f}}}{\rightarrow} \mathrm{RNAP} + \mathrm{mRNA} + \mathrm{DNA},$$

where $\mathrm{RNAP} : \mathrm{DNA_c}$ stands for the closed conformation and $\mathrm{RNAP} : \mathrm{DNA_o}$ for the open conformations of RNAP. Using the steady state approximation and the assumption that the initial step in the initiation of transcription is

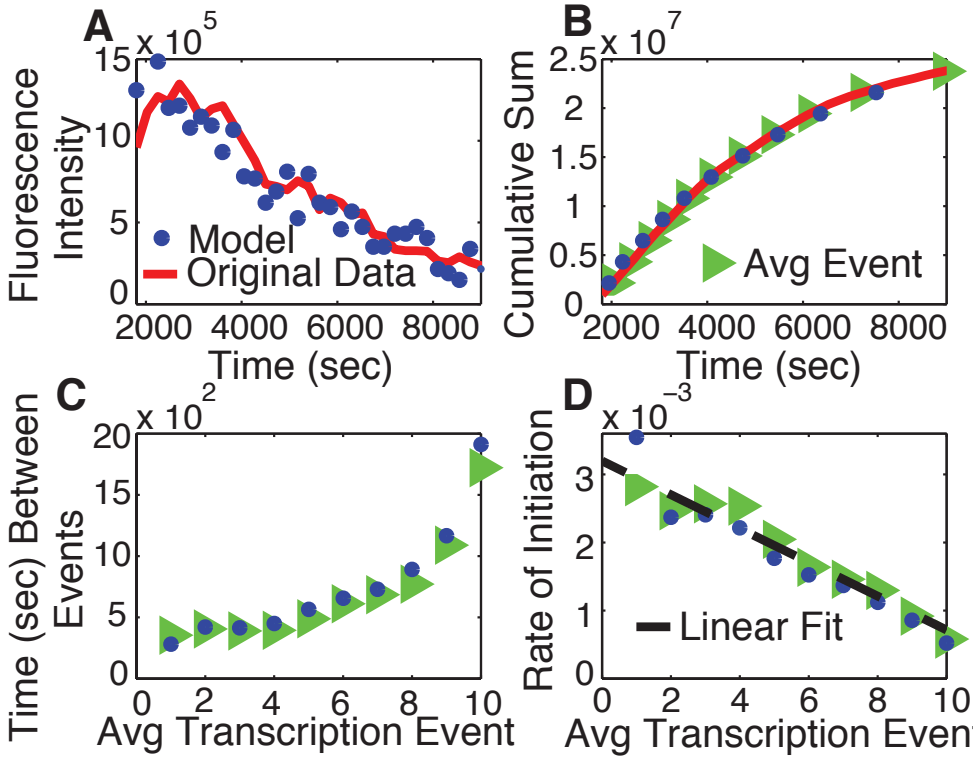

Figure 5.2: (A) Experiment from [82] for T7 RNAP is compared to the results from the model. Where the fluorescence intensity directly corresponds to the number of transcripts produced in the absence of gyrase and the presence of Topo I. (B) The cumulative sum of the data in (A) corresponds directly to the total number of mRNA transcripts produced through time. An average transcription event, see text, determined from the original data and from the model. (C) The time between average transcription events. (D) The initiation rate by transcription event for the experimental data and for the model.

the rate limiting step in transcription [234] the kinetic model above results in the classic Michaelis Menten kinetic equation. Here we should note that the Michaelis Menten equation has also be obtained for the steady state approximation of more complicated kinetic models of transcription initiation [235, 236, 234]. Though, the definitions of the two constants in the Michaelis Menten equation $V_{\max}$ and k_M vary in there definitions for the different kinetic models [235, 236, 234]. Considering the substrate copy number, specific promoter DNA, is much lower than the RNAP in a cell, the production rate of mRNA can be approximated as:

$$V = k_{\text{cat}}/k_M, \text{ where } k_M = \frac{k_{\text{cat}} + k_{\text{off}}}{k_{\text{on}}}. \tag{5.1}$$

The question of interest is how does supercoiling affect these different rates? It has been shown on linear pieces of DNA that the melting of the DNA in the promoter is a minor kinetic barrier [237]. However, positive supercoiling has been shown to increase the melting temperature of DNA and could cause the the stability of the DNA to become influential [238]. Here we make the assumption that supercoiling only influences the k_{off} rate. This assumption is supported by the experimental results that the binding affinity of T7 RNAP for the single stranded promoter sequence is greater and it dissociates slower with the single stranded promoter sequence [239, 237]. Note: we do not rule out the possibility that the stability of the DNA could also influence the other kinetic rates in the model and could vary depending on the particular RNAP. Also, considering there are supercoiling sensitive promoters in *E. coli*, where positive supercoiling can actually increase the transcription rate, more experi-

mental evidence is needed to investigate how the different states of RNAP are influenced by supercoiling.

We then make the assumption that the change in the free energy of the transition state for k_{off} directly depends on the energy needed to melt the DNA of the promoter. In order to provide energetic insights as to how supercoiling would affect the DNA we utilized the statistical mechanical model of supercoiled DNA developed by Sen et al. [240, 241], which was built on the framework provided by Benham [242, 243]. This model showed close agreement with experimental results and was demonstrated over a wide range of supercoiling densities [238]. It should be noted that Benham has also utilized this model to develop a kinetic scheme for reactions with single stranded and double stranded DNA [244]. The free energy of having n melted base pairs, n_j junctions and a certain density of supercoiling, σ is:

$$G(n, n_j, \sigma) = n(\epsilon - T\Delta S) + \frac{n_j}{2}] \times \epsilon_o + G_s(n, \sigma) + K_b T] \times ln[g(n, n_j)],$$

$$G_s(n, \sigma) = \frac{C] \times N(\frac{n}{N} + \sigma)^2}{A^2[1 + (\alpha - 1)\frac{n}{N}]},$$

$$g(n, n_j) = \frac{N(N - n - 1)!(n - 1)!}{(N - n - \frac{n_j}{2})!(n - \frac{n_j}{2})!(\frac{n_j}{2} - 1)!(\frac{n_j}{2})!}.$$

Here, $\Delta S = .024$ kcal/(K mol) is the conformational entropy due to melting a base pair. $\epsilon = 7.9$ kcal/mol and $\epsilon_o = 2.5$ kcal/mol are the base pairing and base stacking energies. The function $g(n, n_j)$ is the degeneracy factor with N total base pairs. $C = 1638\ \frac{\text{kcal}}{\text{mol} \times \text{rad}^2}$, $\alpha = 23.4$, which is dependant upon the bending and the torsional stiffness constants respectively [241] , and $A = 10.4$

base pairs per rad^2, see Methods Section for derivation of supercoiling energy G_s. We use the same parameters as obtained in [241].

Starting with n melted base pairs in a circular DNA loop of N base pairs the probability of having k melted base pairs in the promoter, consisting of Np base pairs, would follow a binomial distribution with a probability of successes equal to n/N. This is only valid when a lower fraction of the DNA is melted n/N$<$ 0.06, otherwise more than one melted base pair will be inside the same melted junction [240]. Here we assume that we are at room temperature and this assumption most likely holds. The probability of having a certain number of melted base pairs given the supercoiling density of the DNA would follow the Boltzmann distribution:

$$P(n|\sigma) = \frac{\sum_{n_j} e^{\frac{-G(n,n_j,\sigma)}{K_b T}}}{\sum_n \sum_{n_j} e^{\frac{-G(n,n_j,\sigma)}{K_b T}}}.$$

Then the change in the free energy barrier of the transition to a melted promoter at a certain σ is:

$$\Delta G(\sigma) = \sum_n \sum_k^{Np} P(n|\sigma)] \times \binom{Np}{k} (\frac{n}{N})^k (1 - \frac{n}{N})^{Np-k} \times ((Np - k)(\epsilon) + \Delta G_s),$$

$$\Delta G_s = G_s(n + Np - k, \sigma) - G_s(n, \sigma).$$

The main effect the supercoiling density has is to alter the probabilities of having a certain amount of DNA melted, which influences the probability of having a melted base pair inside of the promoter region and the amount of energy needed to melt the DNA in the promoter region. Taking the difference

of the transition state relative to the transition state at no supercoiling we obtain:

$$\Delta\Delta G(\sigma) = \Delta G(\sigma) - \Delta G(0). \tag{5.2}$$

Using the parameters obtained in [241], with N=180bp and Np=8bp, we numerically solved for the change in free energy; $\Delta\Delta G(\sigma)$, shown in Fig. 5.3 A. According to transition state theory the rate should depend upon $\Delta\Delta G(\sigma)$ according to: $k(\sigma) = k_{(0)} \times e^{-\frac{\Delta\Delta G(\sigma)}{K_b T}}$, where $k_{(0)}$ is the initial rate with no supercoiling density. Numerical values for $k(\sigma)$ are shown in Fig. 5.3 B. In order to simplify further we approximate $k(\sigma)$ as a simple exponential, demonstrated by the fit to the data generated by the model in Fig. 5.3 B, $k(\sigma) = e^{-w\sigma}$. This would only be true if $\Delta\Delta G(\sigma)$ was linear with σ, which has been the result of papers dating back to 1975 [245]. Plugging in the exponential function into Eq. 1 and flipping the sign in the exponent for k_{off}, where the negative in the exponent disappears because the open complex is trying to maintain the melted DNA, we obtain the following equation for the production of mRNA:

$$V = \frac{k_{\text{cat}} \times k_{\text{on}}}{k(o)_{\text{off}} \times e^{w\sigma} + k_{\text{cat}}} = \frac{k_{\text{on}}}{k' \times e^{w\sigma} + 1}, \tag{5.3}$$

where $k(o)_{\text{off}}$ is the initial off rate with no supercoiling, w is a constant and $k' = k(o)_{\text{off}}/k_{\text{cat}}$. A fit to the experimental data from [82] with the 3 free parameters can be seen in Fig. 5.3 C.

The function above can be approximated using the Taylor series expansion, neglecting the higher terms of $\mathcal{O}(\sigma^2)$, which will be negligible, considering $\sigma << 1$ for $\alpha >> 1$ [241]. We then obtain a linear function for V, which

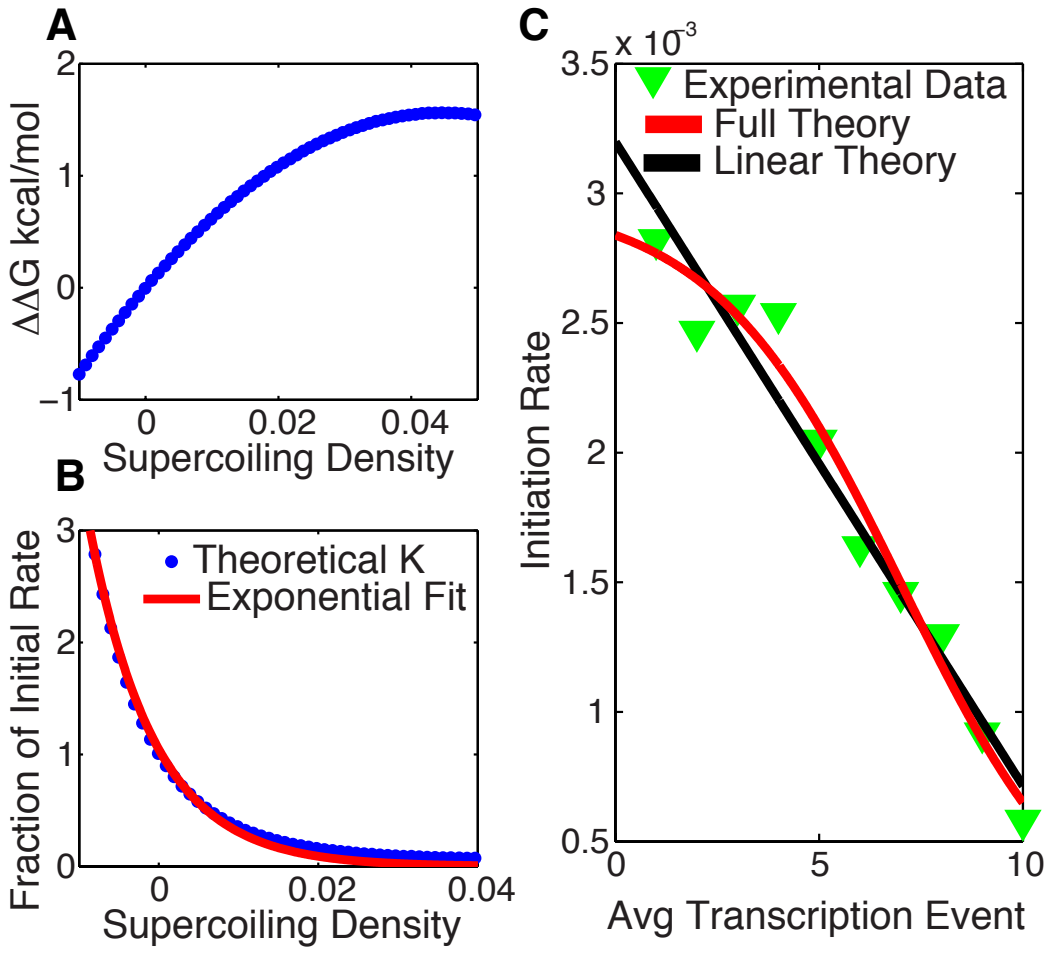

Figure 5.3: (A) The theoretical change in free energy needed to melt the base pairs of the promoter sequence by supercoiling density σ, from Eq. 2. (B) The change in the rate, K, by supercoiling density (dots) and a single exponential fit (line). (C) Transcription initiation rate vs the number of transcription events (green triangles) from experiment [82], the full theory Eq. 3 (red line) and the linear theory Eq. 4 (black line). The full theory had a fit R-square= 0.97 and the linear theory R-square=0.96.

can be simplified further by using the constant $k' = k(o)_{\text{off}}/k_{\text{cat}}$. The y-intercept of the linear fit would then correspond to the production rate at zero supercoiling and the slope would determine how many times the loop of DNA could be transcribed before stalling:

$$V = \frac{k_{\text{on}}}{k' \times e^{w\sigma} + 1} \approx \frac{k_{\text{on}}}{k' + 1} - \frac{k_{\text{on}} \times w \times k' \times \sigma}{(k' + 1)^2}. \tag{5.4}$$

Fig. 5.3C shows a comparison of the full and linear theories. There are most likely many other factors that need to be taken into consideration other than the melting of the DNA when it comes to the production rate of mRNA with supercoiling, though we consider this derivation a starting place to understand how the stability of supercoiled DNA can lead to the initiation rate decrease. We would like to particularly emphasize that this kinetic model is a major simplification, especially considering we do not include abortive initiation and the stability of the DNA could influence other kinetic rates. Also, the system does not necessarily have to start with a supercoiling density of zero, which would then lead to an even greater difference in the free energy to the transition state when supercoiling is accounted for. Due to the easy interpretation of the parameters extracted from the linear fit approximation and the goodness of fit, demonstrating that σ is small, we use the linear model in the remainder of this paper.

5.3 Kinetic model for transcriptional bursting within a supercoiling domain

The accumulation of supercoiling due to transcription is primarily based

on the "twin-supercoiled-domain model" of transcription [227]. Positive supercoiling in different domains builds up due to the absence of gyrase and the presence of Topo 1. Positive supercoiling from transcription has been shown to be a major factor in the build up of supercoiling and positive supercoiling has been shown to have a dramatic effect on the initiation rate of transcription in highly expressed genes [246, 82]. Therefore, it is critical that the buildup and release of positive supercoiling inside specific supercoiling domains be accounted for in modeling these processes. To study the effect of supercoiling on mRNA and protein distributions, we combined our biophysical model, which describes the dependence of transcription initiation on the supercoiling state of the local DNA domain, with a kinetic model of gene expression. We based our kinetic model on a simple burst model of gene expression [230]. Burst models have been frequently used to model stochastic (probabilistic) gene expression [247, 230, 214, 248]. In the burst model, a gene is transcribed to produce mRNA, which is translated to produce protein. Both transcription and translation are first order processes without explicit RNAP or ribosome species. Both mRNA and protein decay also as first order processes. The burst model results in a Poissonian distribution of mRNA molecules and a negative binomial distribution of proteins [249, 217]. Under some conditions, the negative binomial can be approximated by a gamma distribution [230], which is a two parameter distribution relating to the burst frequency and the burst size. Here we present a modified gene expression model such that the transcription rate is linearly dependent on the amount of positive supercoiling that has accumulated in the local DNA domain, in accordance with our simplified biophysical

Table 5.1: Kinetic model for gene expression with local supercoiling effects

Reaction		Propensity
(1)	$DNA + RCoil \rightarrow DNA + PCoil + mRNA$	$a_o \times RCoil$
(2)	$mRNA \rightarrow 0$	$\gamma \times mRNA$
(3)	$mRNA \rightarrow Protein$	$b_o \times \gamma \times mRNA$
(4)	$Protein \rightarrow 0$	$d \times Protein$
(5)	$Gyrase \rightarrow Gyrase'$	$K1$
(6)	$Gyrase' \rightarrow Gyrase$	$K2$
(7)	$Gyrase' + PCoil \rightarrow Gyrase' + RCoil$	$R \times PCoil$

model of transcription initiation. We first define two additional species *RCoil* and *PCoil* that track the amount of "regular", the normal state, and positive supercoiling, respectively, inside the local domain. The sum of these two species is fixed and is denoted by $max(RCoil)$. Production of a transcript converts one *RCoil* into a *PCoil*. Here, we assume an implicit fast relaxation of the corresponding negative supercoiling by Topo 1 [250]. Though, we do not rule out the possibility that other Topoisomerases may be contributing to the dynamics of these systems. The accumulation of *PCoil* linearly decreases the transcription rate of the DNA supercoiling domain according to reaction (1) in Table 5.1. The value of $max(RCoil)$ is therefore equal to the number of times the DNA domain can be transcribed before transcription stalls.

To model the relaxation of the positive supercoiling we introduce a gyrase binding site in the local DNA domain as an additional species that can either be empty *Gyrase* or bound with a gyrase molecule *Gyrase'*. We assume a constant pool of free gyrase such that binding is pseudo first order. Gyrase unbinding follows first order kinetics. When gyrase is bound *PCoil*

is converted to *Rcoil* with a rate constant R that is fast relative to the other rates in the system, such that when gyrase is bound the local domain is effectively completely in the "regular" state. The complete model is shown in Table 5.1 and directly links the accumulation of supercoiling to the number of transcription events that have taken place since the last time that gyrase unbound. To test our kinetic model we simulated the *in vitro* experiment conducted by Chong *et al.* [82] by extracting the needed parameters from the linear fit in Fig 3D. The y-intercept of the linear fit in Fig. 5.2 D corresponds to the maximum initiation rate, $a_o \times max(RCoil) = .0032\text{sec}^{-1}$. The x-intercept corresponds to the number of transcription events that could take place before stalling, $max(RCoil) = 13$, all other rates and species were set to zero. We ran 160 simulations with a single gene in the supercoiling domain the results of the simulated experiment compared to the experimental data can be seen in Fig. 5.2 . In order to take into consideration the stochastic nature of the reactions in Table 5.1 we used the Gillespie algorithm to simulate the system [251]. This was done assuming that the cell is a well stirred environment with no spatial constraints. These simulations where run using the program Lattice Microbes [252] with the rates normalized to the degradation rate of the protein, which is assumed to be on the order of cell division. In the remainder of this study, the rates for the equations in Table 5.1 were $(\gamma = 50, b_o = 2, d = 1, K1 = 10, K2 = 35, R = 1000, max(RCoil) = 4)$ except when specified in different specific situations.The mRNA degradation rate,γ, was chosen so that the lifetime of the molecule would be short, ~ 2 min. The max(RCoil) was set to 4, as it has been previously suggested that 4 rounds

of complete transcription in a supercoiling domain could result in inhibiting transcription, due to the environment of the cell [82]. The rates for promoter strength, a_o, and translation, b_o, can take on a large range of values, we simply choose values that produced means that are physiologically relevant [214].

The range of rates for gyrase in the cell is a matter of debate and likely depend on many different factors. For instance *in vitro* the dissociation constant has been seen to range from 0.2-0.5nM for specific gyrase binding sites to 100nM for not so strong gyrase binding sites [82, 253, 254]. Though, what is taking place *in vivo* adds new factors to the system that need to be taken into consideration, e.g. there are also endogenous inhibitors of gyrase [255, 256]. Furthermore, using the Zero-spike model, discussed later, Chong et al. [82] determined the ratio of the gyrase binding rates to the gyrase unbinding rates for multiple genes through smFISH. We do not believe the values obtained from the Zero-spike model give the exact dissociation constants for gyrase, but we do believe that the fits show there is a wide range in the ratio of these rates. They observed a range of values .1-4.5 demonstrating that the kinetic rates of gyrase vary wildly from gene to gene inside the cell. For our case of K2=35 would correspond to a weak gyrase binding site with a ratio between the two K1/K2=.28. K1 was set to give an average rebinding time of ~ 6 min [82].

5.4 mRNA distributions with supercoiling sensitive transcription

The probability distribution of the number of mRNA molecules per cell

is an important quantity in gene regulation. Many previous studies have utilized smFISH, to quantify the mRNA distribution [225, 82, 214]. To calculate the different distributions of mRNA created by our model, we simulated the equivalent of 4000 cells using the previously specified parameters and ignored the protein part of the distribution, although we keep everything relative to the protein degradation rate of 1. To study how the promoter strength of the gene influences the distribution of mRNA, we simulated one gene with a strong promoter and one gene with a weak promoter by setting a_o=100 and a_o=20, respectively. Choosing these two expression rates results in an mean mRNA copy number of 8 for the highly expressed gene and .4 for the lower expressed gene for the traditional burst model mentioned before. The max(RCoil) was set to 4, as it has been previously suggested that 4 rounds of complete transcription in a supercoiling domain could result in inhibiting transcription [82].

When supercoiling is incorporated into the model, the distribution of mRNA is clearly dependent on promoter strength, in agreement with experiments where genes with higher means also had a higher fano factor [225, 221]. We should mention that mRNA distributions that exhibit bursting can be explained by including extrinsic noise, but here we only investigate the influence of supercoling [221]. When there is only one gene inside the supercoiling domain, the gene with strong promoter maintains a large probability at zero copy number, as seen in Fig. 5.4A. This is qualitatively similar to the distribution of mRNA observed in highly expressed genes fit by the Zero Spike model, discussed below [82], and quite different from the Poisson distribution predicted

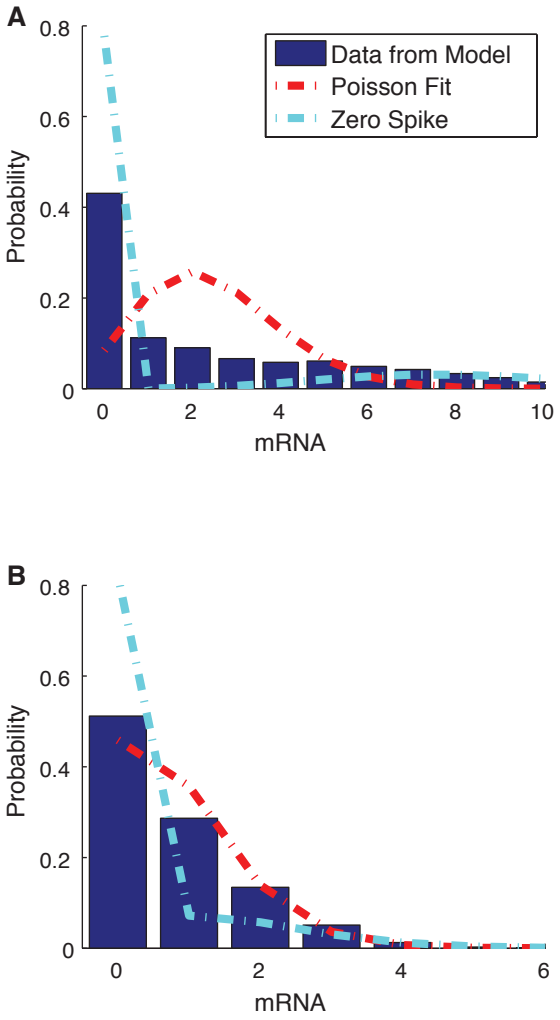

Figure 5.4: (A) The distribution of mRNA for a highly transcribed gene from our model (blue bars), a fit of the simulated data to a Poisson distribution (red line) and a fit to the zero-spike model (cyan line) (B) The same as in (A) for a gene with low expression.

by the burst model.

When the promoter strength of the gene is low, the distribution shows minor deviations from a Poisson distribution, which is predicted by a simple birth and death process (Fig 4B). The degree to which a distribution corresponds to a Poisson distribution can be determined by the Fano factor, the variance divided by the mean. We found that as the binding affinity of gyrase was increased, the Fano factor approached the expected value, while the Fano factor approached one as the initiation rate, a_o, was decreased, Fig. 5.5.

Chong et al. [82] developed a model to obtain a probability distribution of mRNA incorporating the gyrase binding and unbinding, the Poisson with Zero Spike Distribution. This model allows the gene of interest to switch on when gyrase is bound and immediately switches off when gyrase unbinds. However, the accumulation of supercoiling is not linked with the number of transcription events, thus the initiation rate does not decay as supercoiling is accumulated. The distributions predicted by this model deviate substantially from the model proposed here. The distribution obtained by plugging true rates K1, K2, $a_o \times max(Rcoil)$, γ, into the Poisson with Zero Spike Distribution can be seen in Fig. 5.4. The Zero Spike distribution greatly overestimates the probabilities at zero copy number and is not able to correctly predict probabilities of having low copy number of mRNA. The Zero-spike Model converges to our model when K1 is extremely low.

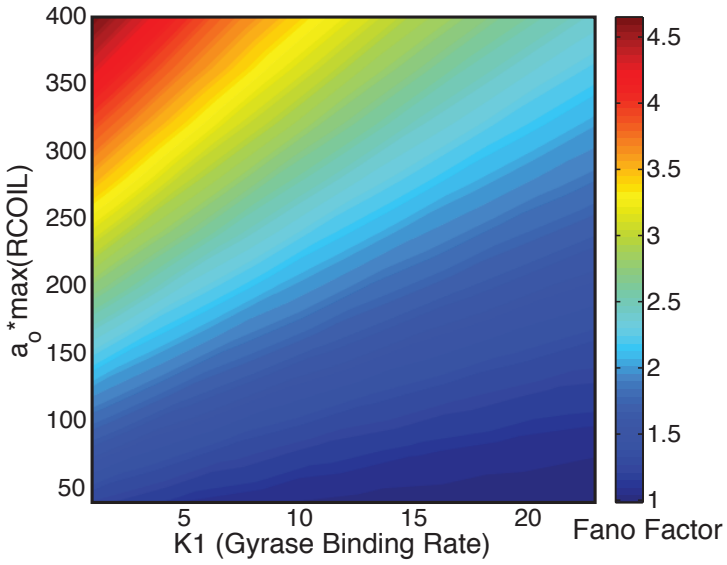

Figure 5.5: The Fano factor, variance/mean, of mRNA of a single gene inside a supercoiling domain with varying initiation rate, a_o, and gyrase binding affinity, "K1".

5.4.1 Protein distributions compared to the burst model of gene expression

We next sought to compare the protein distributions produced by our supercoiling model against those produced by a typical model used for studying stochastic gene expression. The standard burst model of gene expression, shown schematically in Fig. 5.6A, results in a negative binomial distribution of proteins [217]. Under conditions of where decay rate of mRNA is fast relative to that of proteins ($\gamma >> 1$), as is typically the case in bacteria, the discrete negative binomial distribution can be approximated by the gamma distribution [217, 230]: $p(n) = \frac{n^{a-1}e^{-n/b}}{\Gamma(a)b^a}$ where Γ is the gamma function. The rates used in our model would correspond to the gamma distribution with $a = a_o \times max(RCoil)/d$ and $b = b_o \times \gamma/\gamma = b_o$. Fitting protein abundance data to a gamma distribution gives an estimate for two key parameters from the model: a, the burst frequency, and b, the burst size. We wanted to study how well these two parameters could be estimated using a gamma distribution if the underlying data were generated by our kinetic model, which accounts for supercoiling induced transcriptional bursting. We generated simulation data from our supercoiling model using the parameters $a_o = 90$, $max(RCoil) = 4$, $b = 2$, $\gamma = 50$, $K1 = 10$, and $K2 = 35$. The stationary probability distribution for the protein is shown in Fig. 5.6B. The distribution has the typical long tail seen in *E. coli* protein distribution data. We then fit our simulation data to a gamma function. Although the distribution appears to be well-described by a gamma distribution, the a and b parameters from the fit no longer correspond

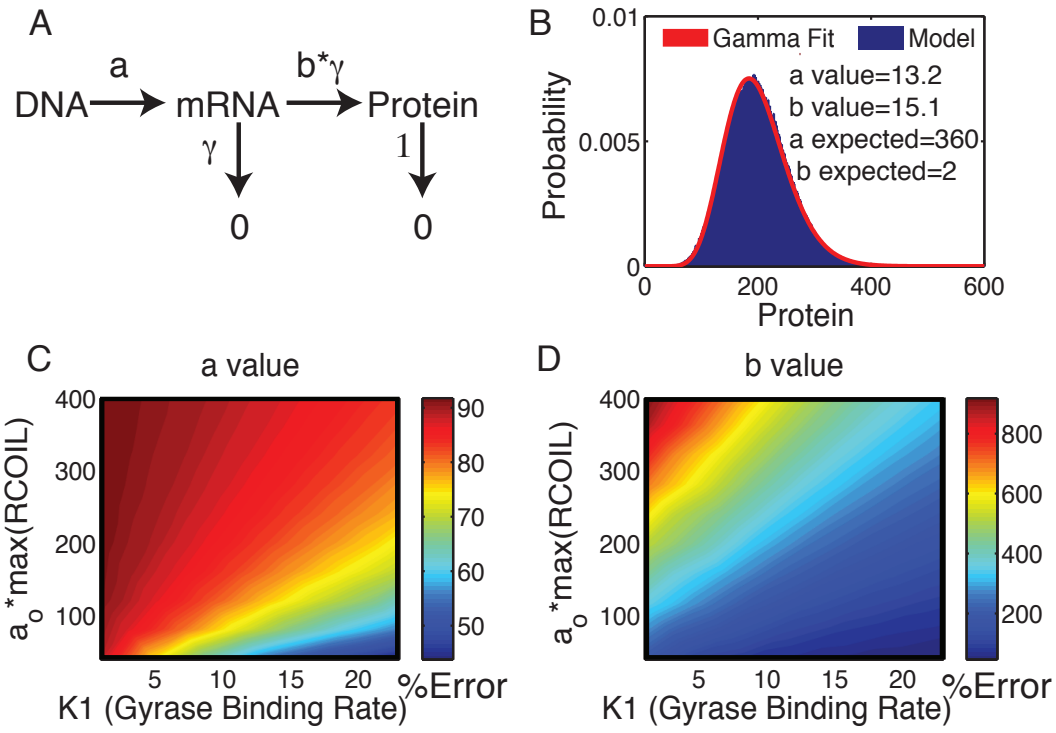

Figure 5.6: (A) The bursting model. (B) The protein distribution generated from our model fit to a Gamma distribution. (Where the probability distribution above is for $K1 = 10$). (c and d) Show the percent error in the a and b values determined by fitting a gamma distribution to the data from the model.

to the model parameters. The estimated a value was 13.2 and the estimated b value was 15.1, which are each approximately an order of magnitude from the correct values.

In a case where the gyrase binding site is predominantly occupied, *i.e.* when $K1$ is high and $K2$ is low, our model converges to the burst model. Likewise, when the initiation rate is comparable to the gyrase binding rate, the two models converge. We were interested in what parameter regions the two models give similar results and performed a parameter scan of $K1$ and a_o

and compared the model parameters versus the gamma fit parameters. Fig. 5.6C+D show the results of the comparison. We consistently find an underestimation of a, which is to be expected as a is an effective burst frequency while a_0 is the basal transcription initiation rate. We also saw a consistent overestimation of b.

5.5 Correlations between genes in supercoiling domains

The supercoiling domains in *E. coli* are thought to be loops roughly 10kb in size [257]. The build-up of positive supercoiling in a local DNA domain affects not only the gene being transcribed, but all other genes in the domain. Here we assume that when a gene in a domain is transcribed the transcription of a neighboring gene does not cancel/enhance the supercoiling generated by its neighbor, though if more than one gene in the supercoiling domain is being transcribed at the same time whether the genes are arranged in a co-directional or divergent pattern will most likely be an important factor.

To study any correlations introduced into gene expression by this coupling, we modified our model to include five genes in a supercoiling domains, when the genes belong to a particular domain we refer to this as a linked domain. The model was modified by having the promoters within the domain firing according to the first reaction in Table 5.1, but with different values for a_o. This means all of the promoters in the domain feel the same supercoiling state, RCoil. We still keep the assumption that the DNA loop can be transcribed a total of 4 times before stalling, as discussed above. Considering there are now

5 genes in the same domain, we set $max(Rcoil)$ to be 20 to allow the whole domain to be transcribed 4 complete times before stalling. The domain has a single gyrase binding site, considering there is roughly one gyrase molecule for each loop in the cell and gyrase is thought to have relatively specific binding sites [258]. Specifically, we simulated two independent supercoiling domains, one which contained some strong promoters and one which contained only weak promoters. The genes that belong to supercoiling domain 1 are genes $1-5$, while the genes that belong to the second independent supercoiling domain are genes $6-10$. The a_o values for the genes from 1 to 10 are 34, 20, 14, 10, 6, 2, 3, 2.4, 0.4, 1.0.

We simulated our two independent supercoiling domains and tracked the expression level of each gene over the course of the simulation. Fig. 5.7A+B show the mean expression levels of the mRNA and protein, respectively. Also shown are the expected values if each gene were in its own supercoiling domain, red. The mRNA and protein levels of relatively weak promoter genes 3-5, which reside in the supercoiling domain with the strong promoter genes, were significantly reduced. The weak promoter genes sharing supercoiling domain 2 were less perturbed.

Next, we calculated the pairwise correlations between mRNA and proteins for all of the genes, $p_{x,y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$. Both mRNA and protein in the same supercoiling domain showed correlation in expression in individual cells and were only correlated with genes in the same loop. The correlation of the genes in individual cells was found to be dependent upon the initiation rate and the binding affinity of gyrase for the supercoiling domain. The effect of expression

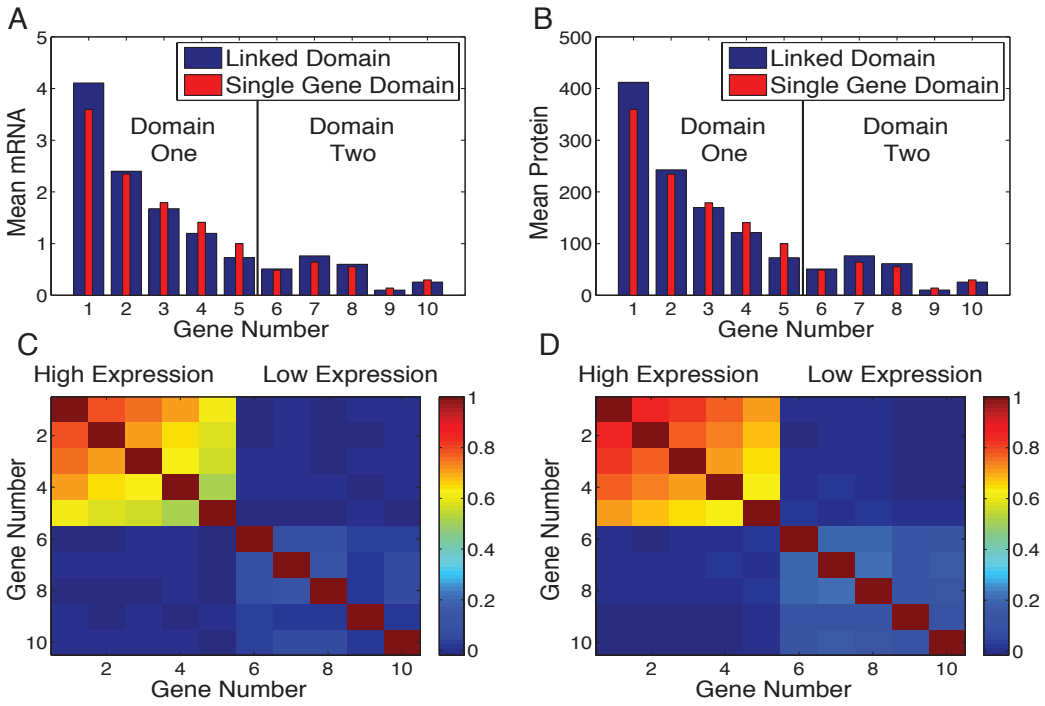

Figure 5.7: (A+B) Genes 1-5 share a supercoiling domain, while genes 6-10 share a supercoiling domain. The red bars indicate the expression level of the gene if it was the only gene in the supercoiling domain while the blue show the means for the linked domain.(B) The correlation for the genes shown in (A+B) in the linked domains.

level on correlation is shown in Fig. 5.7, where the higher expressed genes were significantly correlated. The domain with lower expression genes did not exhibit significant correlation. In both cases the proteins showed slightly higher correlation than the mRNA. A weaker binding affinity of gyrase, K1/K2,

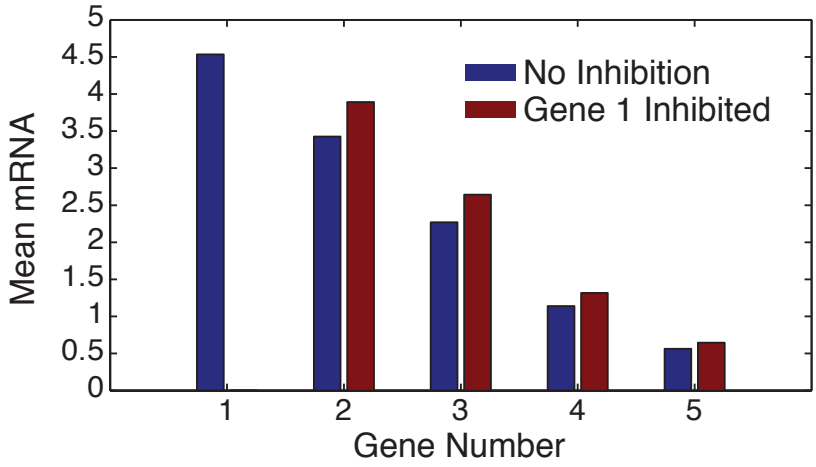

Figure 5.8: The mean mRNA level of a supercoiling domain with all genes expressed are shown in blue, where after gene 1 is inhibited is shown in red.

for the domain also created an increase in the correlation (data not shown).

5.6 Negative regulation within supercoiling domains

We then investigated whether regulating the initiation rate, a_o, of a gene inside the same supercoiling domain would influence the expression level of other genes inside the same domain. The expression level of a domain with five genes was analyzed before, blue bar, and after the inhibition of gene 1, red bars, Fig. 5.8. The inhibition of the highly expressed gene in the domain led to an increase in the expression level of the other genes in the same domain. This could also correspond to positive regulation if a gene were able to repress

itself.

5.7 Conclusion

5.7.1 Supercoiling build-up generates broad mRNA distributions

Our model linking supercoiling to transcription produces mRNA distributions that are noticeably different than those predicted by the burst and two-state models, as shown in Fig. 5.4. Of particular note is the enhanced probability at low (but greater than zero) copy numbers. Because transcription in our model gradually turns off, the distribution is spread over a wider range of low values, rather than being concentrated at the Poissonian distribution for the high state and zero for the low state. Anecdotally this agrees with published mRNA distributions, but a systematic survey must wait until genome-scale mRNA distribution data is available.

It has been shown that the Fano factor of the mRNA distributions is greater than one for highly transcribed genes [214, 221]. This behavior is captured by our model due to the broadening in the mRNA distributions, though we do not rule out the possibility that these distributions can be obtained by including multiple sources of extrinsic noise [221]. It is likely that the bursting behavior seen in mRNA distributions is a contribution of both supercoiling and extrinsic noise.

When analyzing the protein distributions produced by our model, we found that the a and b values deviated greatly from the predicted values of a burst model even though the distribution could be fit by a gamma distribution. If

the fit to a gamma distribution is justified, the a value will correspond to the number of mRNA produced in a proteins lifetime and the b value will correspond to the number of proteins produced per mRNA transcript. Our model shows that when supercoiling is accounted for the a and b values determined from the fit do not correctly represent the underlying processes of the system. Therefore, studies that rely on a physical interpretation of a and b may need to be adjusted.

This does not necessarily mean there is no way to extract the underlying parameters from the distributions of mRNA and protein. Though we consider it to be beyond the scope of this paper, if the gyrase binding constant could be varied in a controlled manner and the distribution of mRNA could be obtained at different gyrase binding rates, then the parameters in this model could potentially be found. Even though there is no unique analytical formula for the distribution produced by this model, the parameters could be found by fitting to simulation data.

5.7.2 Coordination of transcriptional bursts in neighboring genes

Correlation in the transcription of genes in bacteria has been previously reported. In [259], the expression levels of neighboring genes in *E. coli* were shown to be correlated and dependent on supercoiling. The authors concluded that expression levels were directly linked to the distribution of gyrase on the chromosome. Our model predicts this effect; gyrase has differing binding affinities for different supercoiling domains, which effectively controls the the

overall expression level of each gene in the same supercoiling domain. Other authors have reported that expression of an inducible reporter gene represses downstream neighboring genes [260]. Our model also exhibits this effect, as upregulation of one gene (or set of genes) in a supercoling domain reduces expression of other genes in the domain.

Though correlation in the overall expression level of genes has been observed, our model predicts an additional degree of correlation, namely correlation in the transcriptional bursts of all genes in a supercoiling domain. Our model gives rise to correlation of not only protein abundance but also mRNA abundance of neighboring genes in a supercoiling domain, such as shown in Fig. 5.7. This correlation, coupled with the short lifetime of mRNA, means that bursts of mRNA molecules are also correlated in time between genes, *i.e.* neighboring genes are active and inactive in synchrony with each other. Such synchronization could be an important mechanism of transcriptional regulation in bacteria [261]. For instance, genes corresponding to specific functions are known to be located in similar areas on the genome. Synchronous expression of these genes would help to ensure all of the components are produced at the same time.

The correlation of clustered genes has been probed experimentally and clustering was not found to have a significant impact on the expression levels [262]. The authors proposed that any correlations due to gene clusters were washed out by the global extrinsic fluctuations. However, this was only done at a few specific locations in the chromosome and the effect of supercoiling could have been overlooked considering supercoiling not only affects the corre-

lation in overall expression, but influences the correlation though time in the individual cells. In order to observe high correlation we found that the expression level of the genes inside the domain must be high enough to generate enough supercoiling to halt transcription before gyrase binds. We propose that mRNA smFISH of neighboring genes, carefully chosen for expression levels and to lie within a single supercoiling domain, would be an accurate assessment of whether correlation of transcription bursts occurs in *E. coli*.

5.7.3 A structural level of gene regulation

Negative feedback has been proposed to be an important factor in controlling the stochastic nature of the biochemical reactions that take place in gene networks [263, 264]. Given that a transcription event of any gene inside the same supercoiling domain will increase the positive supercoiling felt by all genes inside the domain suggests, there is a higher level of negative regulation at the structural level for bacterial genomes. Every gene within the same domain would essentially negatively regulate the other genes inside that domain. This can be observed in Fig. 5.8, where the expression level of genes inside the same domain increase when a highly expressed gene inside the domain is inhibited. Such an effect could be essential for the proper stoichiometry of the gene products.

Having a built in regulation network inside of the cell would help ensure the proper expression of certain genes without requiring extra energy for the production of transcription regulation factors specific for different genes. In this way, the expression level of genes in the same supercoiling domain would act as an intrinsic regulatory mechanism, providing yet another reason why

the relative ordering of the chromosome is important. Such an effect would have implications in understanding the origin and widespread evolutionary conservation of operonal genome structure in Bacteria and Archaea.

5.8 Methods

5.8.1 Derivation of Transition State Free Energy

This derivation is mostly taken from Sen et al. where an even more in depth explanation of the supercoiling free energy can be found [240, 241]. This derivation treats the DNA molecule as a "homopolynucleotide" in the shape of a circle where there is no base specificity. This DNA molecule has a total of N base pairs and a linking number equal to Lk, where a linking number is the number of base pairs per turn of the B form DNA. When a DNA molecule is supercoiled the linking number will change, ΔLk, which is just the difference between the relaxed linking number and the linking number with supercoiling. The number of base pairs per twist is equal to 10.4 and will be assigned to the letter "A". The twist rate is defined as $2\pi/A$ and the change in the twist rate in the helical regions will be assigned τ_h. Likewise, the twist rate in the melted regions will be τ_c. In order to molecule to be stable the following equation must be obeyed:

$$c_h * \tau_h = c_c * \tau_c$$

Where c_h and c_c are defined as the "torsional stiffness constants." If we then examine the linking number of a partially melted DNA molecule, Lk', with n

melted base pairs Lk' can be represented by the following equation:

$$Lk' = [(N-n)\tau_h + n * \tau_c]/2\pi + (N-n)/A$$

Where the first term is the linking number due to supercoiling and the melted region and the second region is due to the regular melted regions. It is a fundamental that the linking number has to remain constant throughout, Lk'=Lk, considering the molecule is a closed loop. Setting these two equations equal to each other and then solving the previous equations for τ_h and τ_c the following is obtained:

$$\tau_h = 2 * \pi[\Delta Lk + n/A]/[N + (\alpha - 1)n]$$

and

$$\tau_c = 2\pi\alpha[\Delta Lk + n/A]/[N + (\alpha - 1)n] = \alpha\tau_h$$

Where $\alpha = c_h/c_c >> 1$ The supercoiling energy was derived in the studies by Benham and is beyond the scope of this SI for an in depth derivation [242, 243].

$$G_s = \frac{1}{2}(N-n)c_h * \tau_h^2 + \frac{1}{2}n * c_c * \tau_c^2$$

Taking the previous equations for τ_h and τ_c and plugging them into the equation and using $\sigma = \Delta Lk * A/N$, the supercoiling density, gives the free energy of supercoiling for the partially melted DNA molecule.

$$G_s(n,\sigma) = \frac{C * N(\frac{n}{N} + \sigma)^2}{A^2[1 + (\alpha - 1)\frac{n}{N}]}$$

The total free energy of the DNA molecule can then be written as the following:

$$G(n, n_j, \sigma) = n(\epsilon - T\Delta S) + \frac{n_j}{2} * \epsilon_o + \frac{C * N(\frac{n}{N} + \sigma)^2}{A^2[1 + (\alpha - 1)\frac{n}{N}]} + K_b T * lng(n, n_j)$$

$$g(n, n_j) = \frac{N(N - n - 1)!(n - 1)!}{(N - n - \frac{n_j}{2})!(n - \frac{n_j}{2})!(\frac{n_j}{2} - 1)!(\frac{n_j}{2})!}$$

This equation represents having all of the n melted base pairs being distributed into the n_j junctions. But here we are only interested in DNA molecules at room temperature, which corresponds to a small fraction of the molecule being melted. If we look at the average number of n_j for a certain amount of n melted base pairs we see that for small n the number of base pairs in each junction is very small [240, 241]. This allows us to approximate the probability of having a certain number of melted base pairs inside of the promoter to be independent. This is where the binomial distribution results from in the text.

Chapter 6

Bacterial DNA loop formation as a mechanism of "long-range" transcriptional regulation

6.1 Introduction

Cells use a multitude of diverse mechanisms to regulate transcription — constantly adjusting their response to their environment and "neighboring" cellular processes. In all domains of life, many of these regulatory mechanisms have been shown to depend upon the "architecture" of the chromosome, but yet we still do not understand the interplay between the forces directing it's conformation and why certain organizational processes exist.

At the largest scale of organization, Eukaryotic chromosomes have been shown to dictate the positioning of genes — where genes toward the center of the nucleus generally exhibit higher output compared to those found at the periphery [265]. While at a scale of tens to hundreds of kilobases (kb), the chro-

mosomes are organized into topologically associating domains (TADs), where the probability of interaction between intra-domain DNA is much greater than interactions between DNA segments in neighboring domains. These TADs have been shown to essentially limit the contact of intra-domain enhancers to promoters within the same domain coordinating the expression of genes within a domain [266].

Many different factors influence TAD formation within Eukaryotic chromosomes and their precise influence on the chromosome are thought to be dependent upon each other. Transcription itself is often associated with the formation of domains and has been used (with great accuracy) to predict the domain layout along the chromosome [267]. But transcription is not the only driving force behind the formation of these domains (especially in mammals) as argued in the of excellent review by Rowley *et al.* [268]. Two other factors that direct the proper formation of DNA loops are cohesin and CTCF, where it has been shown that cohesin acts as a molecular motor to "extrude DNA loops" [269]. And, while cohesin can start to form a DNA loop independent of CTCF, the CTCF on the DNA does signal to the cohesin to stop extruding, creating a "stable" domain in specific locations.

This brings into question: why expend energy "extruding" the loop in the first place? Is this the most efficient mechanism to form a loop at a particular location or could this be a mechanism of transcriptional bursting or could its role could be restricting the interaction of the enhancer to its given domain

during the formation of a loop [268] — we still do not really know.

Much less is known about bacterial chromosomes. The diversity within bacteria makes generalizations difficult, as the chromosomes can be very different and the "forces" driving their conformation vary in importance. For instance, within *Caulobacter crescentus* the formation of larger chromosome domains are very much dependent upon transcription [270], but domains on a much smaller scale (≈10kb), like that within *E. coli*, are thought to be the result of DNA binding proteins. Interestingly, the locations of these domains within *E. coli* are thought to be highly stochastic (not fixed), which could be one of the reasons why it has been so hard to determine the structure of the *E. coli* chromosome at this scale [80]. The stochastic nature of loop formation was deduced by monitoring the expression of a supercoiling sensitive promoter when the DNA was nicked for a few minimal places along the chromosome, and therefore, whether or not this is a universal rule still needs to be determined.

Still, this important result, along with the pioneering work of Chong *et al.* led to a "convincing" mechanism of transcriptional bursting within *E. coli* [82] (As discussed extensively within the previous chapter) — where it was shown that the buildup of positive supercoiling within a domain can negatively regulate transcription within that domain. To complicate things further, a more recent study showed that the local accumulation of negative supercoiling could lead to premature termination and a slower elongation rate [271]. These re-

sults clearly show how certain properties of the chromosome can have a large influence on the expression of certain genes, especially those which are involved with controlling the propagation of supercoiling — DNA loops [80].

While DNA looping has been extensively investigated regarding an increase in the local concentration of transcription factors and how it results in the cooperativity of transcription factor regulation (where at least 1 of the binding sites are proximal to the promoter region) [272, 203, 273, 274] — very few bacterial studies have investigated the influence of larger loop formation (≈10kb). A recent study produced DNA loops of a comparable size to those naturally found within the *E. coli*, but only developed the system to quantify enhancer promoter interactions and did not investigate if loop formation had any influence on transcription [275]. Interestingly, to which we are aware, no bacterial studies have investigated how "larger" topological domain formation influences the transcription of genes inside and outside of the domain — as is likely the natural case within bacteria.

Here we developed an assay to form a loop (≈6kb) containing multiple genes within the chromosome of *E. coli*. We then quantified the expression levels of the two genes inside of the DNA loop as well as a third gene directly adjacent to the DNA loop boundary with and without loop formation. We found that loop formation led to drastic changes in the expression of each of these genes — where each gene's expression level was dramatically reduced upon loop formation. This was even apparent when the promoters of the

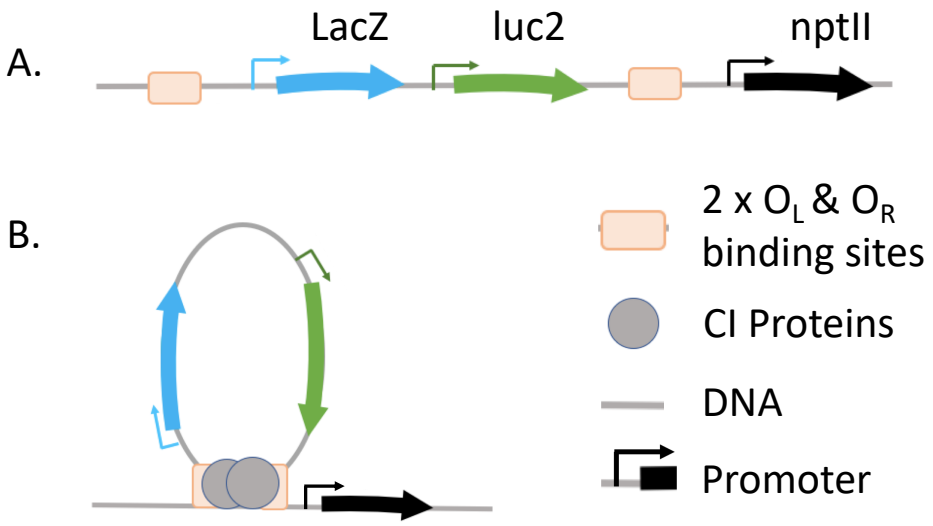

Figure 6.1: A. The looping construct in the unlooped state in the absence of CI. B. The looping construct in the looped state in the presence of CI.

genes were multiple kb away from the DNA loop boundaries, demonstrating that "non-local" DNA architecture has a large role in the regulation of transcription within bacteria.

6.2 Results

6.2.1 DNA loop formation regulates the expression state of "local" genes within and outside of it's boundaries

To directly probe whether DNA loops of a similar size to those naturally found within the *E. coli* chromosome [80], have any influence on the transcriptional

state of "local" genes, we created a construct that forms a DNA loop with the expression of the DNA binding protein CI [203]. Previous studies have shown that the bacteriophage λ repressor protein (CI) can bind to the homologous rightward and leftward operator sites, O_R123 and O_L123, to form a DNA loop [276]. It has also been shown that by increasing the number of CI binding sites, CI is able to form loops larger than 50kb in size [277]. Therefore, to create a construct that could efficiently form a loop, we increased the number of CI binding sites so that there was at least 2xO_L123 and 2xO_R123 and eliminated the natural promoters (P_L, P_R, and P_{RM})(Methods). The construct contained two genes within the CI binding sites, LacZ and luc2 (luciferase), and a third gene (NPTII — Kanamycin resistance) on the outside of the DNA loop, approximately 600bps away from the CI binding sites (Fig. 6.1). The two genes within the DNA loop had the LacUV5 promoter (constitutive expression) [221] and NPTII was driven with the strong EM7 promoter (constitutive expression), as initial constructs with weaker promoters led to problems with selection during insertion of the construct. This large construct (8kb) was inserted into the *E. coli* chromosome (Methods) [278] and loop formation was controlled by the constitutive expression of CI from a plasmid (Methods).

To quantify the different mRNA species within the same cells we performed 3-color single-molecule *in-situ* fluorescence hybridization (smFISH)(Methods) [200] and quantified the mRNA for each gene with and without loop formation. We found that the expression levels for each gene dramatically decreased upon loop formation (Fig. 6.2), suggesting a strong role of DNA loop forma-

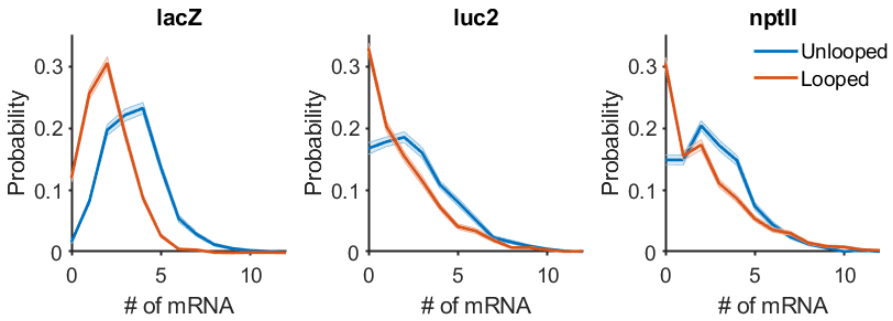

Figure 6.2: The mRNA distributions of each gene with (orange) and without (blue) looping (shaded area is the standard error of mean for each bin determined through bootstrapping).

tion in the regulation of transcription. Upon loop formation we found that not only did the mean expression level of each gene decrease significantly (Fig. 6.3A), but the shape of the mRNA distributions per cell seemed to change in shape. Note: interestingly, this was not just specific for genes within the loop, as the expression level of NPTII was dramatically decreased as well. And, given that the distances (in bp) of each of the genes promoters to the boundaries of the DNA loop (LacZ Promoter>200bp, luc2 Promoter>2000bp, NPTII Promoter>600bp), these results suggest that the topology of the chromosome can have a more "global" "long-range" influence on the transcriptional states of local genes.

6.2.2 DNA loop formation increases the Fano factor while decreasing the mean

Transcription was initially modeled as a simple birth-death process, which

results in a Poisson distribution [279]. But experiments have continually demonstrated that transcription does not follow a simple birth-death process and instead occurs in bursts, increasing the noise of the distribution. Note: the degree of noise within the mRNA distributions is often quantified by the Fano factor: the variance divided by the mean, which equals 1 for a Poisson process.

Transcriptional bursting has been shown to be common within Eukaryotic organisms and Bacteria [280, 82, 223], whose specific fitness advantage is still unclear. The mechanism of transcriptional bursting within bacteria is still a matter of debate — though it is likely a combination of many different mechanisms, whose importance has yet to be determined [281].

One "general" scaling trend for transcriptional bursting within bacteria, is that the higher the mean expression level of the gene, the more noisy the mRNA distribution [221, 223]. The main mechanism behind this scaling relationship has been proposed to be DNA replication and RNAP concentration variation (extrinsic noise) [221]. Supercoiling along with chromosomal architecture has also been suggested to be a mechanism behind transcriptional bursting, whose scaling relationship can be dependent upon a multitude of different complex interactions and does not necessarily need to scale with expression level [81, 82].

To investigate if this trend holds for our system, we calculated the Fano factor for each gene with and without looping. We found that the Fano factor

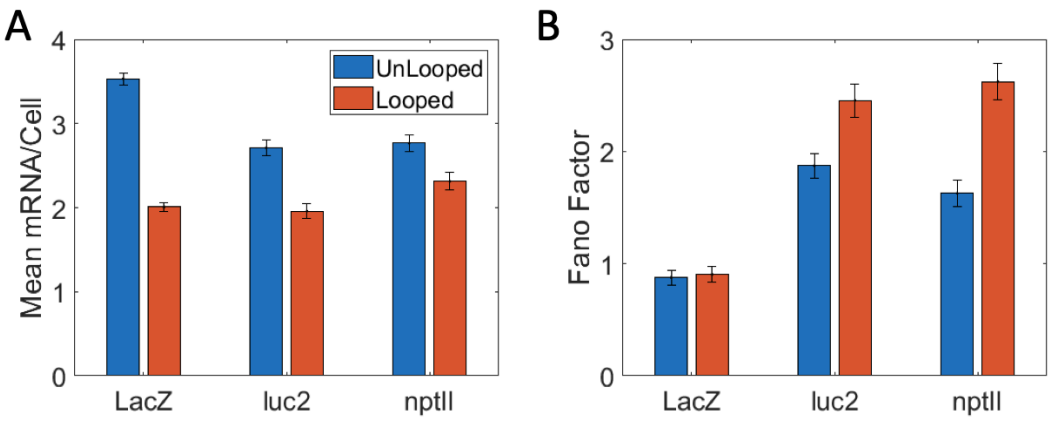

Figure 6.3: A. Mean amount of mRNA per cell with and without looping. B. The Fano factor for each gene with and without looping. Here error bars are two standard errors of the mean, determined through bootsrapping.

significantly increased with loop formation while the mean expression level decreased (Fig. 6.3), supporting the idea that chromosomal architecture (and likely supercoiling) is a main driving force behind transcriptional bursting [82] — as loop formation is able to deviate from the expected trend with loop formation. In summary, these results do not support a general scaling trend for expression level vs. Fano factor, whose dependence is likely governed by a multitude of different processes, including the chromosome state.

6.3 Discussion

Here we performed the first experiments investigating "long-range" transcription regulation due to loop formation within bacteria. To do this we created a synthetic looping construct, that forms loops on a larger scale similar to that found naturally within the *E. coli* chromosome. We were then able to demonstrate that loop formation not only repressed genes within the domain,

but that of a gene whose promoter was located 600 bp outside of the DNA loop.

Considering that CI has been shown to constrain supercoiling [282], the buildup of constrained positive supercoiling within the domain could be the mechanism leading to the repression of LacZ and luc2 [82]. Where with each transcription event positive supercoiling builds up within the domain due to the presence of Topo1 and absence of Gyrase, leading to the inhibition of transcription within the domain until Gyrase binds or the loop becomes "undone". Constrained supercoiling could also be responsible for the decrease in the expression of NPTII, as this could be due to the local buildup of negative supercoiling between the CI binding sites and the NPTII promoter — as negative supercoiling was recently shown to lead to premature termination and a lower elongation rate [271]. Though, all of the complex interactions due to the generation of supercoiling during the elongation step of transcription is difficult to conceptualize and therefore future modeling studies are clearly needed to understand this system.

Another mechanism that could lead to these results is a more "complicated" "entanglement" of the chromosome — where topological domains larger than the distance between binding sites in supercoiled genomes are formed [283]. These more "entangled" structures could occlude/inhibit all of the genes' promoters upon formation, leading to the repression of all three genes. Also, these "entangled" structures would lead to interactions between DNA segments that are "supposed" to be within the loop and the DNA segments on the neighbor-

ing chromosome — which may be one of the reasons for expending energy with cohesin in Eukaryotic organisms — as these "entangled" structures structures would not be formed, restricting enhancer promoter interactions within specific domain. Though, here we should state that the loops within Eukaryotic organisms and Bacteria are on a much larger scale than those investigated *in vitro* [283].

Regardless of the specific mechanism, the work here suggests that the DNA loops formed within the chromosomes of bacteria have a large influence upon the transcriptional state of the cell and therefore the structure of the chromosome must be taken into consideration to understand bacterial transcription regulation.

6.4 Methods

6.4.1 Bacterial strains and plasmids construction

NY45 (codirectional looping construct) was constructed from parental strain XF001 from Fang *et al.* [284]. The chromosomal insertions were made using a Λ RED recombination-based technique for large insertion, Landing Pad technique [285].

Landing Pad donor plasmid containing the construct used within this work (pSHAY3a) was constructed through In-Fusion cloning (Takara). Through

electroporation, pSHAY3a was transformed into the XF001 strain containing pTKRED (From Dr. Thomas E. Kuhlman, [285]). Several small cultures of the XF001 strain containing pTKRED and pSHAY3a were grown in 10μL of Lysogeny Broth (LB) at 30C for 4 hours. These small cultures were supplemented with 200μL of EZRDM rich media with Streptomycin and grown again at 30C for 4 hours. Finally, the Landing Pad reaction was induced by adding 2mL EZRDM media with 2mM IPTG, 0.4% Arabinose, and 10μg/mL Kanamycin to each culture and the cultures were allowed to grow overnight at 30C. The following morning, the cultures were diluted 1:1000 in LB media, 100μL of each was plated on LB plates with 50 μg mL^{-1} and the plates were incubated overnight at 37C.

Colonies were tested for insertion through PCR using primers along the construct and on the native sections of the chromosome near the insertion. Once successful colonies were identified, the strains were cured of the pTKRED plasmid by growing cultures in 3 mL of LB with 10 μg mL^{-1} Kanamycin at 42C for 7 hour and diluting 1:3000 into 3 mL of LB with 10μg mL^{-1} Kanamycin to grow at 42C overnight.

The resulting strain, NY45, was then transformed with a plasmid containing pACL18, which contains an endogenously expressed CI Λ-repressor gene, to produce strain NY50. As a control, NY45 was also transformed with a plasmid containing pACL18noCI, which was constructed by removing the gene for CI Λ-repressor from pACL18 through In-Fusion cloning (Takara), to produce

strain NY51. Strains NY50 and NY51 were used for the smFISH experiments within this work.

6.4.2 Single Molecule Fluorescent in situ Hybridization (smFISH)

All DNA oligos used for smFISH were ordered from Biosearch Tech: LacZ (Gene 1) transcripts were labeled with 48 oligonucleotides conjugated with Fluorescein Dye, Luc2 (Gene 2) transcripts were labeled with 48 oligonucleotides conjugated with CAL Fluor Red 590, and NPTII (Gene 3) transcripts were labeled with 30 oligonucleotides conjugated with Quasar 670 Dye. See Tabel 6.1 and Table 6.2 for specifics of sequence.

Cultures of NY50 and NY51 were grown overnight in LB media with 25 μg mL^{-1} Chloramphenicol at a temperature of 30C. Control cultures of XF001 were grown overnight in 3mL LB media with 6μL Tetracycline at a temperature of 30C. The following morning, cultures were diluted 1:100 and allowed to grow until the cultures reached an OD of 0.5. Cultures were fixed in a solution of 3.7% (vol/vol) formaldehyde for 30 minutes at room temp. As a wash step, samples were spun down and resuspended in 1x PBS 2 times. The fixed cells were then permeabilized in 70% (vol/vol) ethanol for 1 hour. Samples were spun down and resuspended in 1x PBS 1 time. The samples were then spun down again, resuspended in hybridization buffer (40% (wt/vol) formamide, 2x SSC) with a 1x oligonucleotide probes and left to incubate overnight. The following day, samples were spun down and resuspended in 1x PBS 2 times

and stored on ice. This procedure was adapted from [200].

Fixed and labeled cells were imaged using brightfield conditions and cycles of excitation at 488 nm, 561 nm (Coherent, sapphire), and 647 nm (Coherent, obis). Each field of view was imaged at four z planes evenly spaced by 250nm. The maximum projection of the four images for each field of view was used to count the mRNA in each cell. The total fluorescence intensity of each spot in the images was divided by the intensity of one mRNA and rounded to calculate the amount of mRNA?s in that spot [200]. Image analysis was done using custom MATLAB scripts.

6.4.3 Probes

smFISH Probes 1 of 2		
LacZ Probes	luc2 Probes	NPTII Probes
Fluorescein Dye	CAL Fluor Red 590	Quasar 670 Dye
gattaagttgggtaacgcca	ggccttttttgatgtttttt	tcttgttcaatggccgatc
aaaggggatgtgctgcaag	aacgtttcattgctttgtgc	ggagaacctgcgtgcaatc
aactgttgggaagggcgatc	gatgtgtgcgtcggtgaatg	aatagcctctccacccaag
tgaggggacgacgacagtat	ttctgcgtaggtgatgttta	ctgttgtgcccagtcatag
tgtagatgggcgcatcgtaa	cagacgtacggacatttcga	gcatcagagcagccgattg
gtaatgggataggtcacgtt	cgtaacgtttcattgcttct	cgctgacagccggaacacg
gggaacaaacggcggattga	atacggtggttggtgttcag	acaaaaagaaccgggcgcc
atgtgagcgagtaacaaccc	ggagttttcggagcatacta	tgcagttcattcagggcac
tagccagctttcatcaacat	cagtaccggcatgaagaact	cacagctgcgcaaggaacg
aaaataattcgcgtctggcc	tactgctacgccgatgaaca	gcttcagtgacaacgtcga
ttgcaccacagatgaaacgc	gttcgttgtagatgtcgttt	gcacttcgcccaatagcag
ttcagacggcaaacgactgt	atgttcatggagttcagcag	agatgacaggagatcctgc
cgcgtaaaaatgcgctcagg	atacgaatactacggtcggc	ctttctcggcaggagcaag
tcctgatcttccagataact	tacgttcaggattttctgca	attgcatcagccatgatgg
gagacgtcacggaaaatgcc	tggatgatcggcagtttttt	cggatcaagcgtatgcagc
tgtgtagtcggtttatgcag	gttttggagtccatgatgat	tggtcgaatgggcaggtag
ggcaacatggaaatcgctga	cgaaggtgtacatggactgg	cgatgcgatgtttcgcttg
cgcggctgaaatcatcatta	aagtcgtattcgttgaagcc	cttccatccgagtacgtgc
cacatctgaacttcagcctc	ttttgtcacggtcgaaggat	atcctgatcgacaagaccg
gtaggtagtcacgcaactcg	gagttcatgatcagtgcgat	tgatgctcttcgtccagat
caccctgccataaagaaact	agaaacgtacgcatgcggta	ttgagcctggcgaacagtt
ctcatcgataatttcaccgc	atgatctggttgccgaagat	gggtcacgacgagatcatc
agacgtagtgtgacgcgatc	cggtactacggacaggattg	tgatattcggcaagcaggc
cacagtttcgggttttcgac	ggtacatcagtactacacgg	aaaagcggccattttccac
gcacgatagagattcgggat	aacgcaggaacagttcttct	cggccacagtcgatgaatc
caggcttctgcttcaatcag	gactggattttgtagtcctg	ccaacgctatgtcctgata
agcagcagaccattttcaat	aaggagaacagggtcggtac	gccaagctcttcagcaata
taacgcctcgaatcagcaac	tatttgtcgatcagggtgga	aagcacgaggaagcggtca
tgaccatgcagaggatgatg	atttcgtgcaggttggacag	gaatcgggagcggcgatac
ttcatcagcaggatatcctg	tacttctttggacagcggtg	gatagaaggcgatgcgctg

Table 6.1: Table (1 of 2) showing the oligo sequences for probes used in smFISH

smFISH Probes 2 of 2		
LacZ Probes	luc2 Probes	NPTII Probes
Fluorescein Dye	CAL Fluor Red 590	Quasar 670 Dye
cacggcgttaaagttgttct	ggaaacgttttgctactgct	NAN
agcggatggttcggataatg	tgatcaggattgcggaggtg	NAN
gggtttcaatattggcttca	aacggtactactttgcctac	NAN
gatcatcggtcagacgattc	gtctactactttgcttcga	NAN
gatcacactcgggtgattac	tttacgtagccggacatgat	NAN
atacagcgcgtcgtgattag	ctttgtcgatcagtgcgttg	NAN
ggatcgacagatttgatcca	gaagaagtgttcgtcttcgt	NAN
cgcgtacatcgggcaaataa	agggatttcagacggtctac	NAN
aagccatttttgatggacc	ctggtagcctttgtatttga	NAN
tattcgcaaaggatcagcgg	caggatggattccagttctg	NAN
aaaccgccaagactgttacc	tcgaagatgttcgggtgctg	NAN
aaacgcctgccagtatttag	gtgttccagtactactactg	NAN
cctgtaaacggggatactga	ctttttcggtcatggttttg	NAN
gttgccgttttcatcatatt	gatgctacgtagtctacgat	NAN
ttcggcgtatcgccaaaatc	agtttttttgcggtggttac	NAN
cgttcatacagaactggcga	ttcggtacttcgtctacgaa	NAN
aaactgctgctggtgttttg	tttcacggattttacgtgcg	NAN
tttgcccggataaacggaac	cctttttttgctttgatcag	NAN

Table 6.2: Table (2 of 2) showing the oligo sequences for probes used in smFISH

Chapter 7

Spatial organization of RNA polymerase and its relationship with transcription in *E. coli*

[1]

7.1 Background

Prokaryotes are traditionally viewed as bags of freely diffusing enzymes. This view is rapidly changing. New studies now document that bacterial cells possess a remarkable degree of spatial organization of cellular components and activities without the use of membranes, offering a level of functionality and regulation previously underappreciated [286, 287, 165, 288]. In both *E. coli* and *B. subtilis* cells grown in rich media, RNA polymerase (RNAP), the only enzyme responsible for all RNA transcription, was found to form dense foci instead of distributing homogenously within the cell [289, 290]. Because the majority of cellular RNAP is dedicated to ribosomal RNA (rRNA) synthesis

[1] Weng X*, Bohrer CH*, Bettridge K, Lagda AC, Cagliero C, Jin DJ, Xiao J. Spatial organization of RNA polymerase and its relationship with transcription in Escherichia coli. Proceedings of the National Academy of Sciences. 2019 Oct 1;116(40):20115-23.

in fast-growing cells [291], a transcription factory model was proposed [292]. This model suggests that dense RNAP foci are clusters of hundreds of RNAP molecules actively engaged in rRNA transcription and that their formation is driven by active rRNA synthesis in fast-growing cells under optimal growth conditions (such as LB, 37°C) [289, 293, 292]. This prokaryotic transcription factory model is reminiscent of the RNAP I transcription factory model in eukaryotic cells, in which RNAP I form concentrated, membrane-free condensates in the nucleolus for rRNA transcription [294, 295].

Understanding how and why RNAP is spatially organized in bacterial cells is important as this information could provide new insights into the mechanism of transcription regulation in a complex, heterogeneous cellular environment. For example, in eukaryotic cells, it was suggested that RNAP clusters might represent pre-formed transcription complexes that are "poised" ready for rapid transcription induction [296, 297, 298, 299]. In bacterial cells, such a role has not been demonstrated, but studies have shown that there are typically higher levels of RNAP association at promoter and promoter-like sequences than within coding sequences [300, 301, 302, 303, 304, 305]. However, partially due to technical limitations in dissecting the subcellular organizations of small bacterial cells, these possibilities remain unexamined. In particular, despite a number of recent studies that extensively investigated the distribution and characteristics of RNAP clusters in *E. coli* [89, 137, 40], whether RNAP clusters observed in fast-growing cells are indeed active in rRNA transcription, and whether RNAP clusters only form in the presence of active rRNA transcription, have not been directly examined. Previous studies have shown that

treating cells with rifampicin, a global transcription inhibitor [306], largely abolished the appearance of RNAP foci [293, 307, 40]. However, it remains unclear whether this change was due to diminished rRNA transcription activity, or the associated nucleoid structural changes under the condition of global transcription inhibition [308, 14].

Here we characterized the spatial organization and transcription activity of RNAP under different conditions using quantitative superresolution imaging in *E. coli* cells. We demonstrate that there is an rRNA transcription activity-independent spatial organization of RNAP in *E. coli*, and that the underlying nucleoid structure plays important roles in organizing RNAP clusters.

7.2 RNAP forms distinct clusters in cells growing in rich defined medium

To investigate the spatial organization of RNAP in *E. coli*, we used a strain in which the chromosomal rpoC gene encoding for the β' subunit of RNAP was replaced by a photoactivatable fluorescent protein gene fusion, rpoC-PAmCherry [137, 40, 309]. We verified that the resulting RpoC-PAmCherry fusion protein was expressed in full-length (Methods, Fig. 7.2A), was incorporated efficiently into the RNAP core enzyme complex (Methods, Fig. 7.2B) and supported wild-type (WT)-like cell growth as the sole cellular source of the β' subunit (Methods, Fig. 7.2C). Therefore, the spatial distribution and dynamics of the RpoC-PAmCherry fusion protein should be representative of the native RNAP core or holoenzyme. In the text below, we refer to this fusion

protein as RNAP-PAmCherry for simplicity.

Using RNAP-PAmCherry, we performed single-molecule localization-based superresolution imaging [5] on exponentially growing live cells in EZ Rich Defined Media (EZRDM) at room temperature (25 °C, cell doubling time = 73 ± 1 min, hereafter termed as the rich medium growth condition, Methods, Fig. 7.2C) with a measured two-dimensional (2D) spatial resolution of 50 - 60 nm (Methods, Fig. 7.3). We observed clustered distributions of RNAP-PAmCherry in individual cells (Fig. 7.1A). These clusters were distinct but less punctate compared to what were observed under faster growth conditions (EZRDM 37°C and LB at 37°C, Methods, Fig. 7.4A), consistent with what was reported previously [137, 89, 40]. The averaged cellular distribution of all RNAP localizations displayed a two-lobed pattern with a clear cleft in the middle (Fig. 7.1B), similar to that of the nucleoid imaged using three-dimensional (3D) structured illumination superresolution microscopy (SIM, Methods, Fig. 7.5A). In contrast, free PAmCherry molecules and PAmCherry fused to a non-specific DNA binding protein HU both exhibited a significantly more homogenous distribution in cells (Methods, Fig. 7.6A). Using a truly monomeric mEos3.2-fused RNAP fusion protein [33], we further verified that the clustered distribution was not due to the weak dimerization property of PAmCherry (Methods, Fig. 7.6C, D). Additionally, we developed a stringent algorithm to eliminate false clusters caused by repeated localizations of same molecules due to the blinking of fluorophores [312, 313], and still observed a clustered RNAP distribution (Methods, Fig. 7.6E). Note that all the data used in this work was processed using the algorithm to eliminate repeated lo-

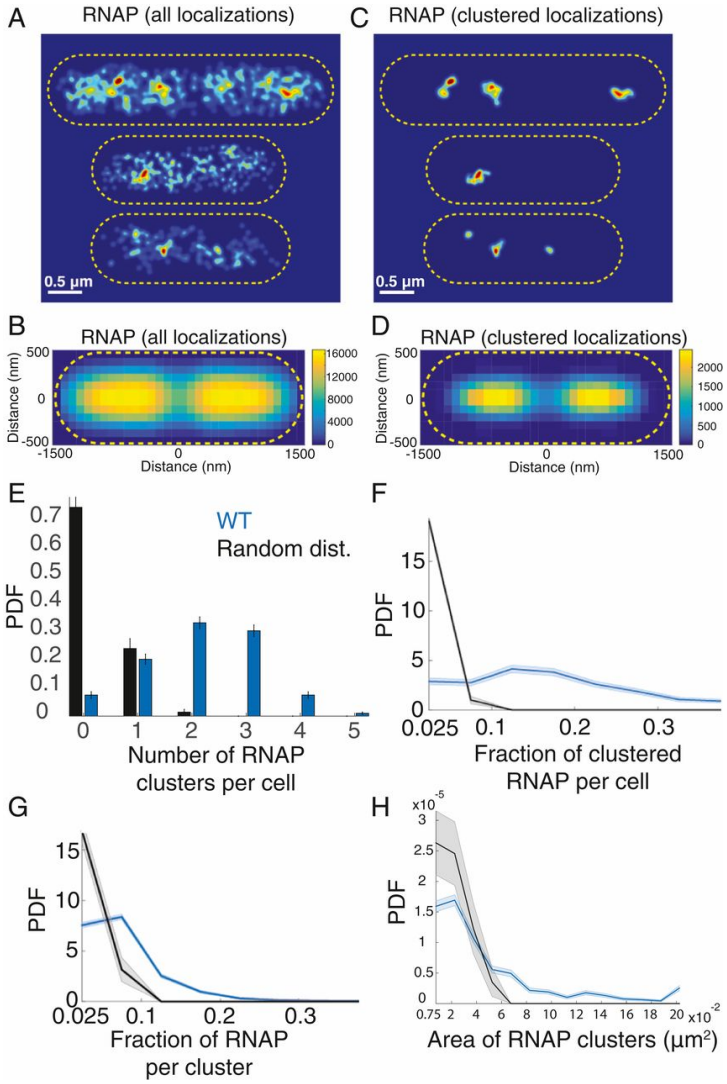

Figure 7.1: Quantitative characterization of RNAP clusters in live *E. coli* cells. (See following page for detailed caption)

Figure 7.1: Quantitative characterization of RNAP clusters in live *E. coli* cells. (A) Representative superresolution images of RNAP (RpoC-PAmCherry) in three cells under the rich medium growth condition. Cell outlines are indicated in yellow dashed lines. Scale bar, 0.5 μm. (B) Two-dimensional (2D) histogram of all RNAP localizations in a standard 3 μm x 1 μm cell under the rich medium growth condition. Because of the symmetry of the cell shape in both long and short axes, we calculated the absolute displacement of each RNAP localization to the center of the cell, normalized its long axis displacement to the standard cell length, and duplicated the quartile cell histogram along both the long and short axes to produce a full-sized 2D histogram of RNAP distribution. The bin size of the 2D histogram is 100 x 100 nm. The color bar indicated localization numbers used in each bin. A total number of 564615 localizations of 664 cells were used to construct the 2D histogram. (C) Identification and isolation of RNAP clusters using a tree-clustering algorithm. RNAP clusters identified in the three cells in (A) are shown as examples. (D) 2D histograms of RNAP localizations in clusters as plotted in (B), a total number of 39438 localizations of 1385 RNAP clusters were used. (E) Distribution of the number of RNAP clusters per cell (blue bars), PDF is probability density function. The mean is 2.13 $\pm$ 0.05 RNAP clusters per cell, $\mu \pm SE$, n = 664 cells. (F) Distribution of the fraction of clustered RNAP per cell. The mean is 0.16 $\pm$ 0.005, $\mu \pm SE$, n = 664 cells. (G) Distribution of fraction of RNAP localizations per cluster. The mean is 0.076 $\pm$ 0.001, $\mu \pm SE$, n = 1385 clusters. (H) Distribution of the area of RNAP clusters. The mean for the radius is 129 $\pm$ 25 nm $\mu \pm SE$, n = 1385 clusters (assuming circularly shaped clusters). In all the graphs from (E to H), the blue curves are the experimentally measured distributions, and the black curves are those calculated from simulated random distributions using the same number of RNAP localizations in the same cell volume for all the cells. Error bars or shaded areas are standard errors calculated from bootstrapping. The average value of each graph is also summarized in Methods, Fig. 7.22 .

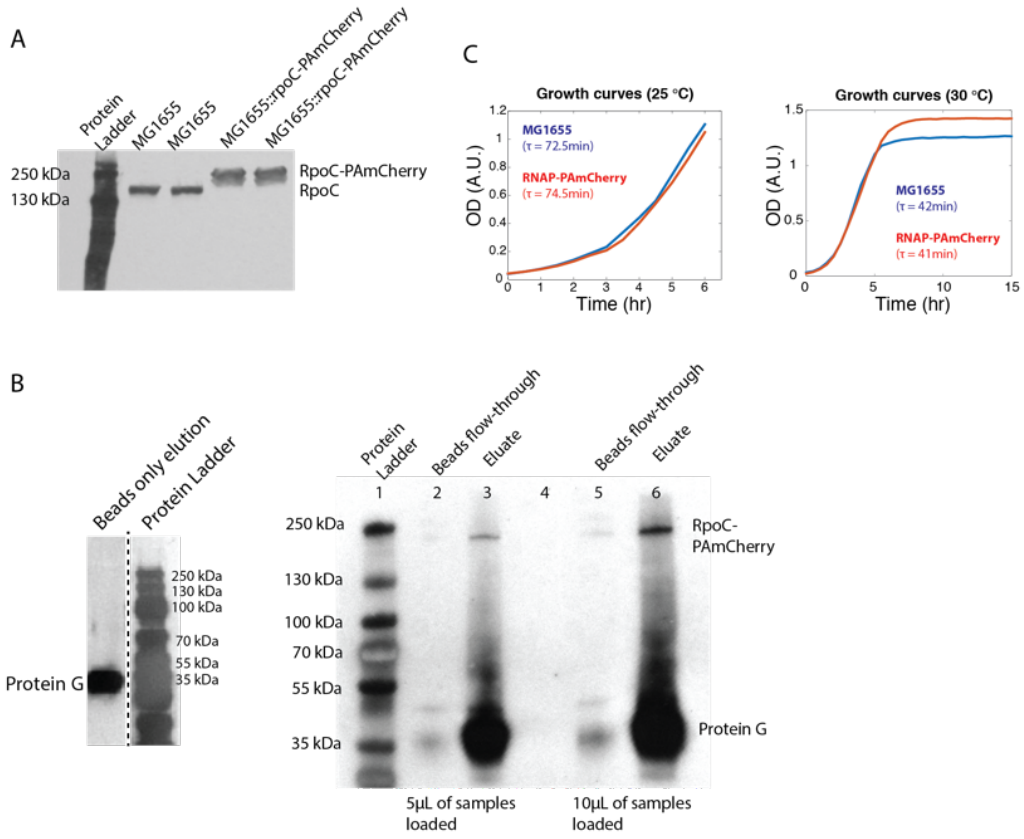

Figure 7.2: RpoC-PAmCherry was expressed in full-length and supported normal cell growth as the sole copy of cellular RpoC. (A) Western blot showed that RpoC-PAmCherry was expressed at the correct molecular size as a full-length fusion (detected by α-RpoC). The MG1655 strain is the wild-type (WT) parental strain of the RpoC-PAmCherry strain. (B) Co-immunoprecipitation of RNAP core and holoenzyme from *E. coli* cell lysates that expressed RpoC-PAmCherry using saturating amount of RpoB antibody conjugated protein G agarose beads and detected using mCherry antibody. Lane 1: protein molecular weight marker. Lane 2 and 3: beads flow-through and eluate, 5 μl loading volume each. Lane 4, blank. Lane 5 and 6, beads flow-through and eluate, 10 μl loading volume each. The majority (88%) of RpoC-PAmCherry was detected in the beads eluate but not the beads flow-through indicating that almost all RpoC-PAmCherry is incorporated inside RNAP core or holoenzyme. (C) Growth curves showed no significant difference in cell doubling times between MG1655 and RNAP-PAmCherry strains under the rich medium growth condition. Growth curves are shown for both RT (25°C) and 30°C.

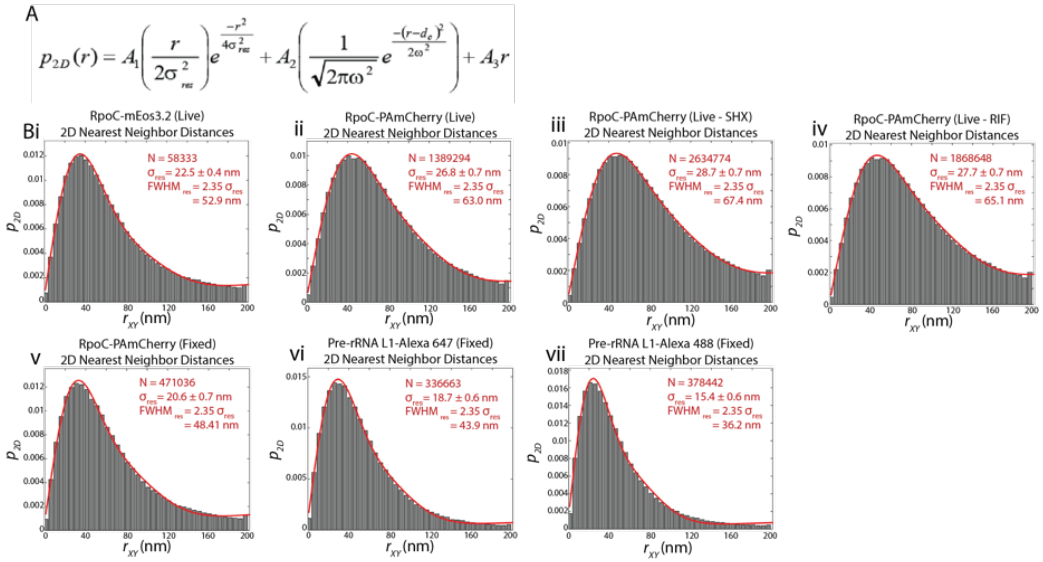

Figure 7.3: Measurement of spatial resolution in single-molecule localization based superresolution imaging. (A) Equation describing the two-dimensional (2D) distribution (p_{2D}) of distances (r) between the nearest neighbors in adjacent frames of localization data with the corresponding localization precision σ_{res}. This equation [310, 311, 167] accounts for the 2D distance distribution expected from repeat localizations of the same molecule (1st term) and the possibility that one molecule's nearest neighbor in the adjacent frame may be another molecule (2nd and 3rd terms described by the Gaussian parameters ω and d_c, and the weight factors A1, A2, and A3). (Bi to Bvii) 2D distance distributions $P_{2D}(r)$ (gray bars) between nearest neighbor localizations in adjacent frames for all listed conditions used in the work and the corresponding fit (red) using the equation in (A). The number of data points N, the fit Gaussian localization precision σ_{res}, and the corresponding spatial resolution $FWHM_{res}$ [144] are listed in each graph.

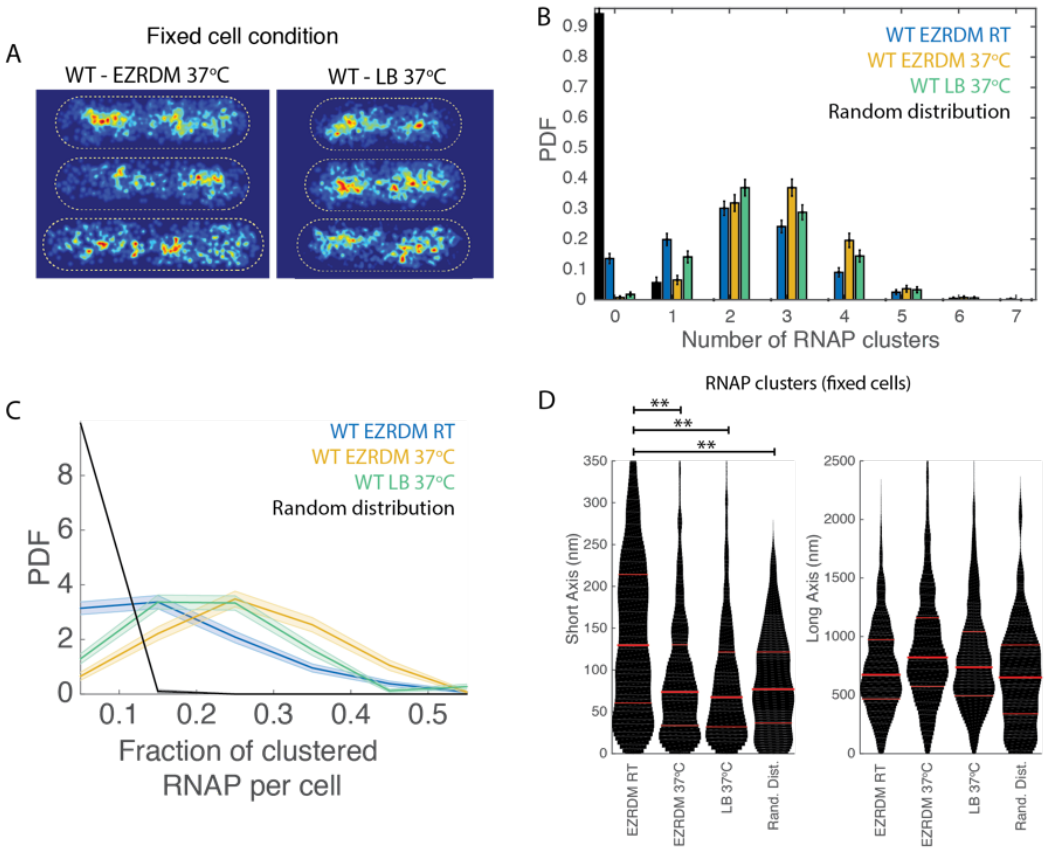

Figure 7.4: *E. coli* RNAP showed a more punctate clustered distribution under faster cell growth conditions. (See the following page for additional details.)

Figure 7.4: *E. coli* RNAP showed a more punctate clustered distribution under faster cell growth conditions. (A) Example superresolution images of RNAP-PAmCherry under EZRDM 37°C and LB 37°C growth conditions in fixed cells. (B) Comparison of the number of RNAP clusters per cell distribution between different growth conditions. PDF is probability density function. The black histogram was that obtained using simulated random distribution. The average number of clusters per cell detected for EZRDM 37°C is 2.8 ± 0.06 ($\mu \pm SE$, n = 276 cells), and for LB 37°C is 2.5 ± 0.04 ($\mu \pm SE$, n = 333 cells). (C) Comparison of the fraction of clustered RNAP per cell distribution between different growth conditions. The average fraction of clustered RNAP per cell detected for EZRDM 37°C is 0.26 ± 0.006 ($\mu \pm SE$, n = 276 cells), and for LB 37°C is 0.22 ± 0.007 ($\mu \pm SE$, n = 333 cells). (D) Averaged cellular positioning of the centroids of RNAP clusters along the short and long axes of cells respectively under different growth conditions. All data were from fixed cell experiments and all cells' sizes are normalized to a standard cell size of 1 μm x 3 μm. Cell center is defined as (0,0). Means are shown as middle red lines in the distributions, with 25th and 75th percentiles shown as flanking red lines. In the main text, these distances were converted back to 3D radial distances by dividing a projection factor 0.64 [169]. The statistical significance of the comparisons with the EZRDM RT condition (indicated by asterisks **: $p < 0.001$) are listed in Methods, Fig. 7.23.

calizations. Therefore, we concluded that the clustered distribution of RNAP-PAmCherry reflected the property of RNAP and not the fusion fluorescent protein or imaging conditions.

To characterize RNAP clusters quantitatively, we performed a density-based threshold analysis to isolate individual RNAP clusters (Fig. 7.1C, 7.22, 7.3, Methods). The averaged cellular distribution of RNAP localizations inside clusters also showed a similar, nucleoid-like pattern (Fig. 7.1D), but was more toward the center of the nucleoid compared to that of all RNAP localizations (Fig. 7.1B). In cells under faster growth conditions (37°C LB or EZRDM) , RNAP clusters were even more inwardly distributed along the short axis of the cell (Methods, Fig. 7.4B-D). In contrast, when compared to the clusters of free PAmCherry and HU, RNAP clusters were located closer to the periphery of the cell (Methods, Fig. 7.7). On average, we detected 2 to 3 dense RNAP clusters per cell (Fig. 7.1E). These clusters contained 16% of total detected cellular RNAP-PAmCherry molecules (Fig. 7.1F), corresponding to approximately 350 RNAP molecules per cluster, given an average of 5000 molecules of RNAP per cell (Fig. 7.1G, Methods) [314, 315]. Note, the number of RNAP molecules within the cells from the fast growth conditions are much higher, leading to a greater number of RNAP per cluster. On average these clusters occupied an area equivalent of that of a circle with a radius of 130 nm (Fig. 7.1H). These cluster properties were significantly different from that of free PAmCherry molecules, what would be expected from a completely random distribution pattern (Fig. 7.1E-H, black curves, Methods, Fig. 7.6, 7.7, 7.23) or cluster properties from that of the non-specific DNA binding HU (Methods,

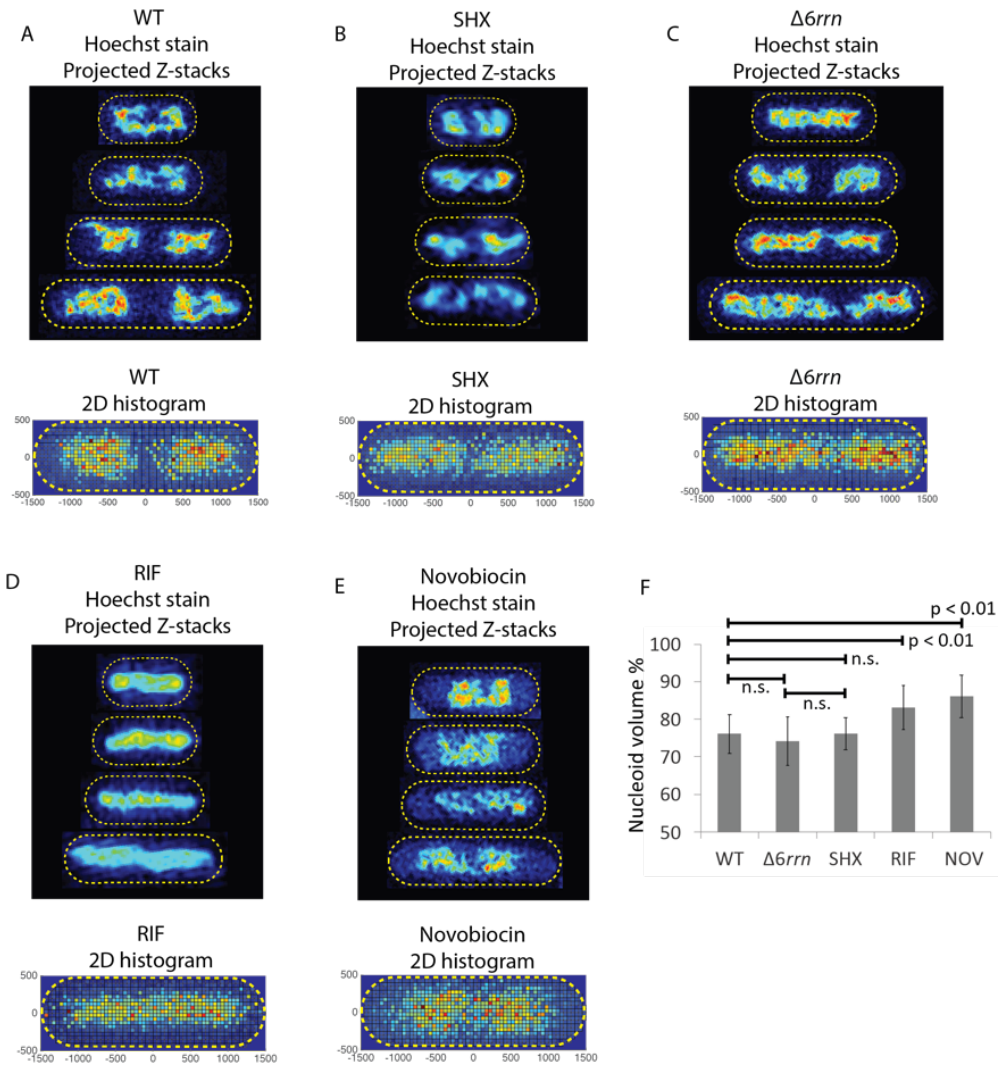

Figure 7.5: Three-dimensional (3D) structured illumination microscopy (SIM) images and analysis of fixed *E. coli* nucleoids stained with Hoechst dye under different experimental conditions. Representative cells are shown for the rich medium growth (A), serine hydroxamate-treated (B), Δ6rrn (C), rifampicin-treated (D) and Novobiocin-treated (E) conditions. Each example cell is shown as projected Z-stacks (maximum intensity projection of 8 x 125-nm interval Z-slices). The bottom panel for each condition shows the overlaid 2D intensity histogram of 15 representative cells for each condition, normalized in a standard 3 μm x 1 μm cell. (F) Average percentages of nucleoid volume over total cell volume calculated from SIM images for all cells under different conditions. The error bars represent standard deviations. P-values were calculated using two-tailed students t-test and a p-value < 0.01 was considered significant.

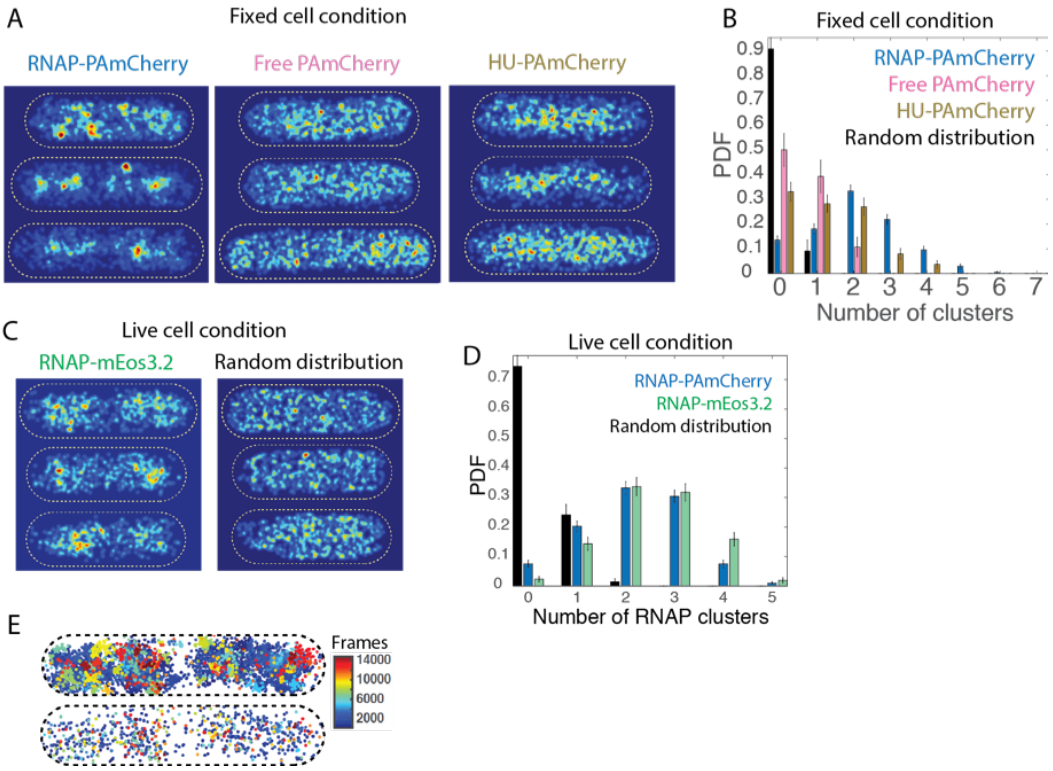

Figure 7.6: *E. coli* RNAP showed a clustered distribution independent of fluorescent protein fusions or fluorophore blinking. (See the following page for additional details)

Figure 7.6: *E. coli* RNAP showed a clustered distribution independent of fluorescent protein fusions or fluorophore blinking. (A) Example superresolution images of RNAP-PAmCherry, free PAmCherry and HU-PAmCherry in fixed cells. (B) Comparison of the number of RNAP clusters per cell distributions for RpoC-PAmCherry, free PAmCherry and HU-PAmCherry in the fixed cell conditions. PDF is probability density function. The black histogram was that obtained using simulated random distribution. The average number of clusters per cell detected for free PAmCherry is 0.61 ± 0.09 ($\mu \pm SE$, n = 56 cells), and for HU-PAmCherry is 1.2 ± 0.08 ($\mu \pm SE$, n = 163 cells). The statistical significance of the comparisons with RNAP-PAmCherry was provided in Methods, Fig. 7.23. (C) Example superresolution images of RNAP tagged with the monomeric fluorescent protein RpoC-mEos3.2 [33] in live cells and the simulated images of random distributions using the same number of localizations of each cell. It was not possible to obtain experimental superresolution images of free mEos3.2 in live cells due to the rapid diffusion of free mEos3.2 molecules. (D) Comparison of the number of RNAP clusters per cell distribution between RpoC-PAmCherry and RpoC-mEos3.2, The black histogram was that obtained using simulated random distribution. (E) Blinking correction using a density correction algorithm [313]. The top is a cell with all detected RNAP localizations prior to blinking correction; the bottom is the same cell after blinking correction. The color bar indicates frame numbers. All the data in this work has been corrected for fluorophore blinking.

Fig. 7.6A, B, 7.7, 7.23), therefore confirming the clustering of RNAP in live *E. coli* cells under the rich medium growth condition.

7.3 RNAP clusters colocalize with nascent rRNA synthesis sites in cells under the rich medium growth condition

Previous studies proposed that RNAP clusters are actively engaged in rRNA transcription but no direct evidence have been provided [292]. To examine this hypothesis, we probed the colocalization of RNAP clusters with nascent, or newly synthesized, rRNAs. We used a highly efficient fluorescence in-situ hybridization (FISH) probe conjugated with Alexa Fluor 488 or 647 (Methods, Fig. 7.8 and 7.9A) to target the 5' leader region of the 16S precursor rRNA (pre-rRNA, Fig. 7.11A), which is absent from the mature 16S RNA inside the ribosome [195]. The 5' leader degraded rapidly with a half-life of 130 sec after being processed (Methods, Fig. 7.9B); therefore, the FISH probe only identifies newly synthesized pre-rRNA. Using two-color superresolution imaging of pre-rRNA and RNAP-PAmCherry in fixed cells, we observed clear spot-like foci of pre-rRNA fluorescence signal with a spatial resolution of 40 nm (Fig. 7.11B-middle panel, Methods, Fig. 7.3). On average we detected 4 pre-rRNA clusters per cell containing more than 60% of total cellular rRNA localizations (Fig. 7.11C-F, Methods, Fig. 7.24, 7.3). Furthermore, we observed qualitatively that RNAP-PAmCherry clusters predominately coincided with these pre-rRNA clusters (Fig. 7.11B-right panel). To quantify the extent of spatial colocalization, we calculated the fraction of RNAP clusters that had

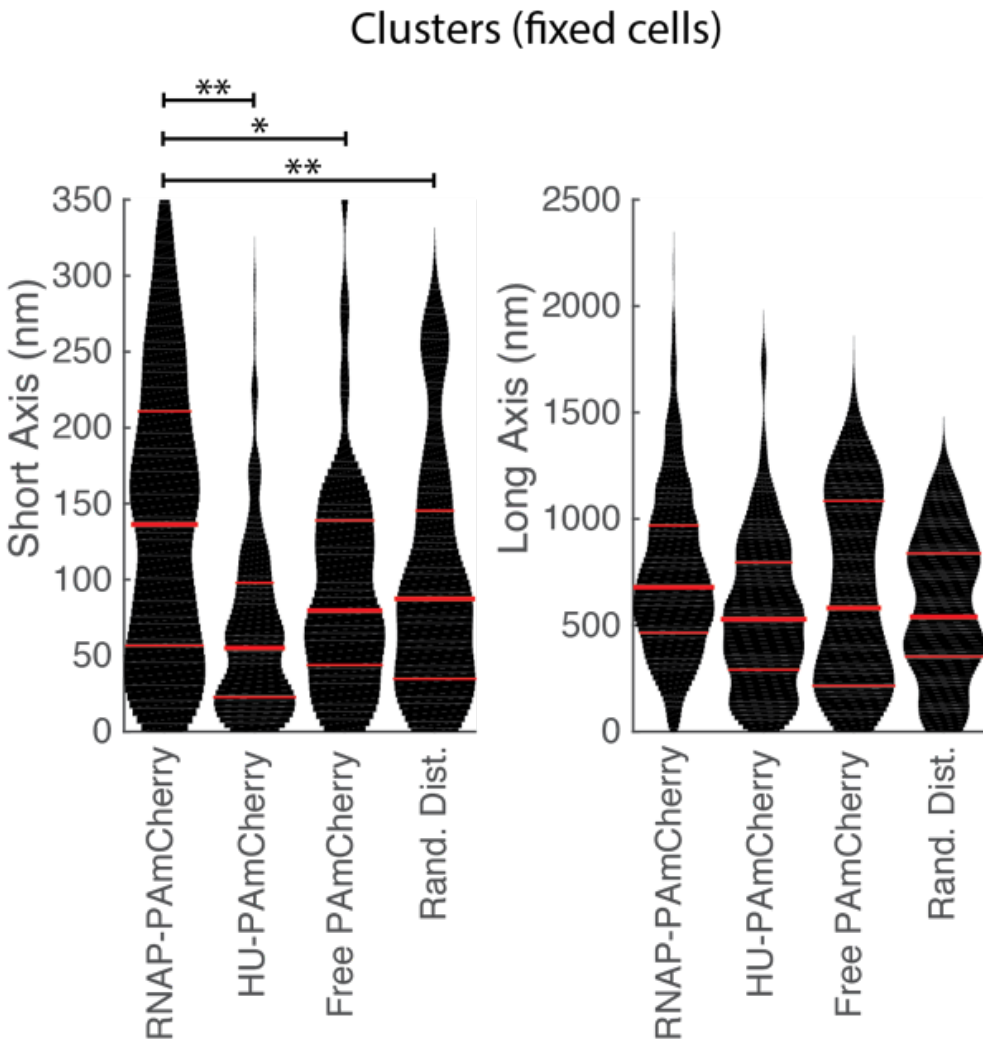

Figure 7.7: Averaged cellular positioning of the centroids of RNAP, HU and free PAmCherry clusters along the long and short axes of cells. All data were from fixed cell experiments and all cells' sizes are normalized to a standard cell size of 1 μm x 3 μm. Cell center is defined as (0,0). Means are shown as middle red lines in the distributions, with 25th and 75th percentiles shown as flanking red lines. In the main text, these distances were converted back to 3D radial distances by dividing a projection factor 0.64 [169]. Asterisks indicate *: $p < 0.01$, **: $p < 0.001$.

any molecule localizing to any molecule of a pre-rRNA cluster within a radius ranging from 50 nm to 250 nm (Fig. 7.11G, blue curve, Methods, Fig. 7.10). We then compared the colocalization curve with the expected background level calculated by randomizing the positions of RNAP clusters in the same cells (Fig. 7.11G, black curve). We found that at all radii there were substantially higher fractions of RNAP clusters colocalizing with pre-rRNA clusters than that of the background level. For example, 83% +- 2% RNAP clusters (n = 404 RNAP clusters) had at least one pre-rRNA cluster within a radius of 50 nm (Methods, Fig. 7.25). Given the significantly improved spatial resolution afforded by superresolution imaging, the high colocalization levels we observed between RNAP clusters and pre-rRNA clusters at a resolution limit (40 – 60 nm) comparable to the molecular size of RNAP molecules (20 nm [316]) suggesting that the majority of RNAP clusters were active in rRNA synthesis under the rich medium growth condition.

7.4 RNAP forms clusters in the absence of high levels of rRNA synthesis

Previous work has proposed that rRNA synthesis is the major driving force for the formation of RNAP clusters [293, 292, 307, 317]. To test this hypothesis, we treated cells with serine hydroxamate (SHX) to perturb rRNA transcription and subsequently observed the spatial organization of RNAP. Serine hydroxamate binds to seryl-tRNA synthetase, induces the stringent response, and inhibits rRNA synthesis [318, 319, 320]. We observed a dramatic reduction in total rRNA synthesis in SHX-treated cells as expected (3% com-

A L1 probe sequence

```
5′(Alexa Fluor-488)TGCCCACACAGATTGTCTGATAAATTGTTAAAGAGCAGTGCCGCTTCGCT

5′(Alexa Fluor-647)TGCCCACACAGATTGTCTGATAAATTGTTAAAGAGCAGTGCCGCTTCGCT
```

B 5′Leader target sequence alignment (non-coding strand)

```
              3′ TCGCTTCGCCGTGACGAGAAATTGTTAAATAGTCTGTTAGACACACCCGT - Alexa Fluor 5′
                 ||||||||||||||||||||||||||||||||||||||||||||||||||
rrnA          5′ AGCGAAGCGGCACTGCTCTTTAACAATTTATCAGACAATCTGTGTGGGCA
rrnB             AGCGAAGCGGCACTGCTCTTTAACAATTTATCAGACAATCTGTGTGGGCA
rrnC             AGCAAAGCGGCACTGCTCTTTAACAATTTATCAGACAATCTGTGTGGGCA
rrnD             GCCGCGTCGCAACTGCTCTTTAACAATTTATCAGACAATCTGTGTGGGCA
rrnE             AGCGCGTCGCAACTGCTCTTTAACAATTTATCAGACAATCTGTGTGGGCA
rrnG             AGCGAAGCGGCACTGCTCTTTAACAATTTATCAGACAATCTGTGTGGGCA
rrnH             GCCGCGTCGCAACTGCTCTTTAACAATTTATCAGACAATCTGTGTGGGCA
                   *    **  **************************************
```

Figure 7.8: L1 probe sequence design (A) and alignment with the targeting regions of the seven pre-rRNA leaders (B). Two L1 probes with the same sequence but two different dye labels (Alexa Fluor 488 and Alexa Fluor 647) were used in this study. The L1 probe is designed to match perfectly the rrnA, B, and G pre-rRNA leader sequence. Starred bases in (B) are completely conserved across all of the seven rrn operons' leader sequences.

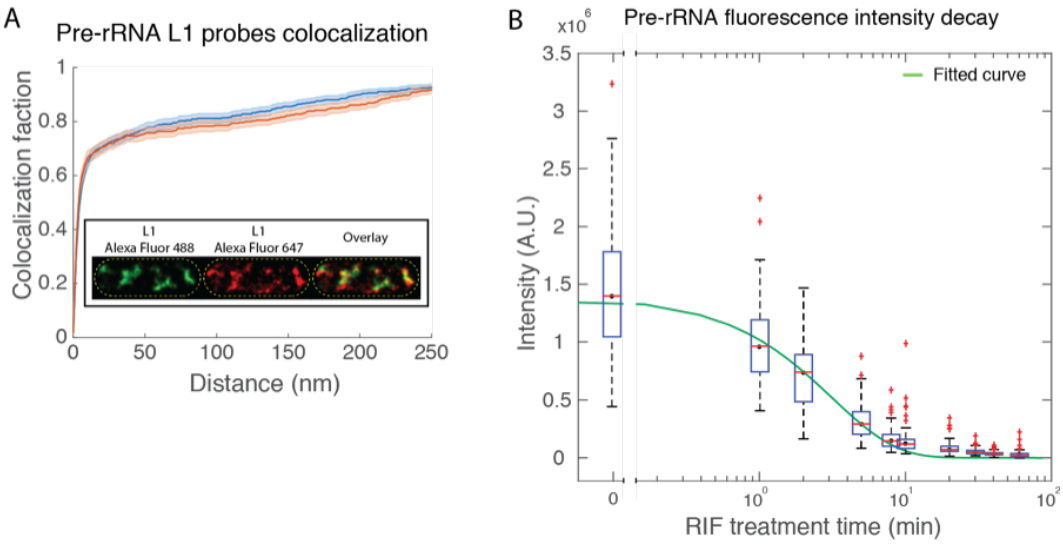

Figure 7.9: Characterization of FISH probe L1 for pre-rRNA detection. (A) Detection efficiency measurement of the L1 probe. Two L1 probes with the same sequence but different dye labels L1-Alexa Fluor 488 and L1-Alexa Fluor 647 as those in Methods, Fig. 7.8A were used to hybridize with the same cells and imaged in two-color superresolution (inset). The high colocalization fractions of one probe to the other (red and blue curves) indicated high detection efficiencies of pre-rRNA clusters using either dye-labeled probe. The detection efficiency was estimated to be 80% for either probe at the distance threshold of 50-nm. (B) Rapid decay of pre-rRNA FISH signal after the inhibition of global transcription using rifampicin. Integrated ensemble pre-rRNA FISH fluorescence intensities of individual cells are plotted at each time point after rifampicin treatment (100 μg ml^{-1}), and fit with a single exponential (green) with a decay rate constant of 0.32 min-1, corresponding to a half-time of 130 sec. The distribution of fluorescence at each time plot is plotted as box plots, with the population mean as the red line, and boxed region as the 25th and 75th percentiles, outlier points defined as data points exceeding 2.7 standard deviations of the distribution are marked in red.

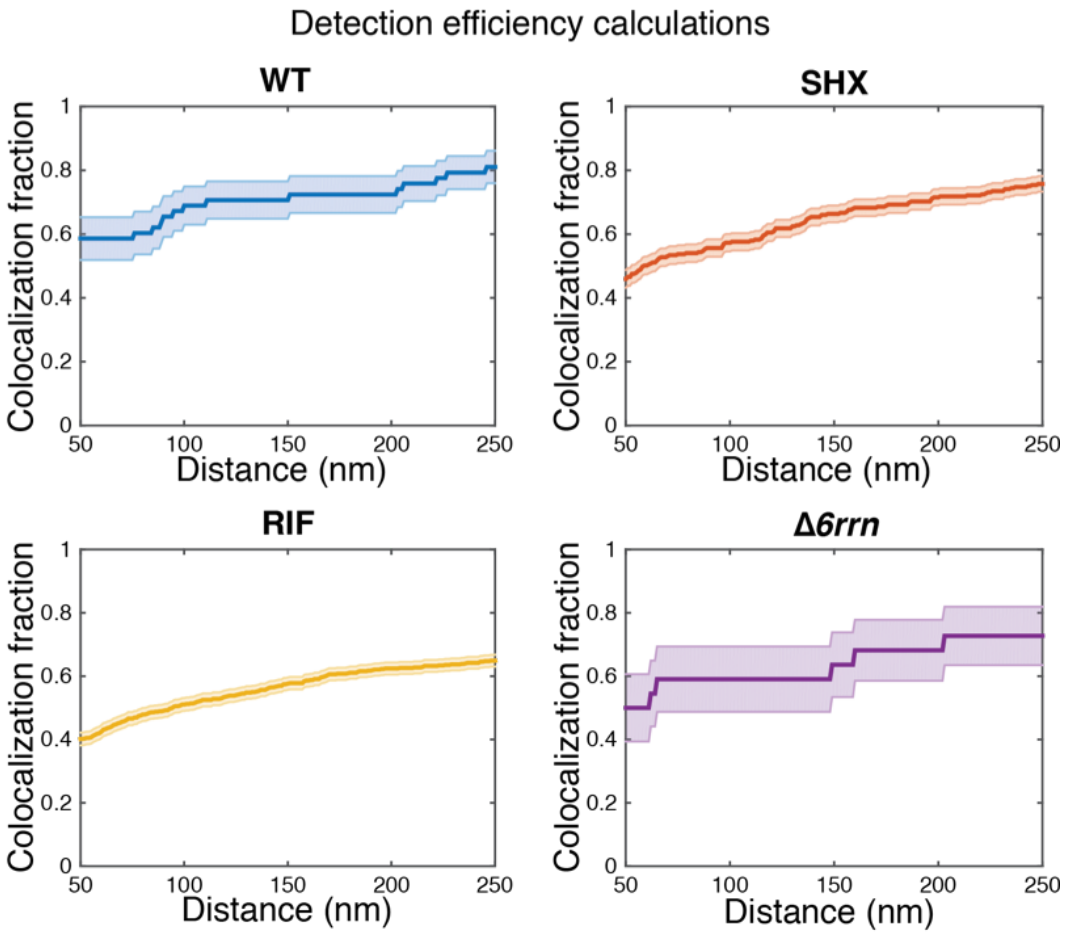

Figure 7.10: Detection efficiency of RNAP clusters at different distances, shown as colocalization fractions with itself for all live cell imaging conditions (WT, SHX, RIF, and Δ6rrn strain). See Methods for details of the calculation.

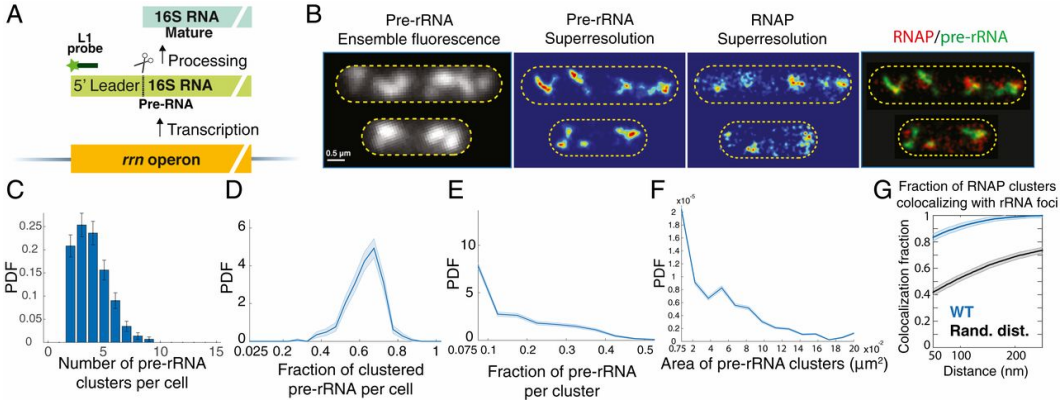

Figure 7.11: RNAP clusters colocalized with nascent pre-rRNA clusters under the rich medium growth condition (See following page for additional details.

Figure 7.11: (A) Schematics of pre-rRNA detection. The dye-labeled L1 probe binds to the 5' leader sequence of 16S rRNA that is cleaved off from mature 16S rRNA and rapidly degrades. (B) Left: ensemble pre-rRNA FISH images of cells (outlined in yellow) under the rich medium growth condition. Scale bar, 0.5 μm. Middle: representative pre-rRNA FISH superresolution images of the two cells. Right: representative two-color superresolution images of RNAP-PAmCherry (red) and pre-rRNA FISH (green) of the two cells in the middle. (C) Distribution of the number of pre-rRNA clusters per cell. The mean is 3.86 $\pm$ 0.09, $\mu \pm SE$, n = 288 cells. (D) Distribution of fraction of clustered pre-rRNA localizations per cell, PDF is probability density function. The mean is 0.63 $\pm$ 0.005, $\mu \pm SE$, n = 288 cells. (E) Distribution of fraction of pre-rRNA localizations per cluster. The mean is 0.16 $\pm$ 0.004, $\mu \pm SE$, n = 1086 pre-rRNA clusters. (F) Distribution of the area of pre-rRNA clusters. The mean for the radius is 127 $\pm$ 22 nm, $\mu \pm SE$, n = 1086 pre-rRNA clusters. The average value of each graph is summarized in Methods, Fig. 7.24. (G) The fraction of RNAP clusters colocalizing with pre-rRNA clusters at different distances from 50 to 250 nm (blue curve). The black curve is the simulated colocalization faction of RNAP clusters with pre-rRNA clusters when the spatial distribution of RNAP clusters was randomized in the same cells, and hence represented the basal level of colocalization due to chance. The plotted colocalization fraction is corrected for detection efficiency of pre-rRNA clusters (Methods, Fig. 7.9A, 7.10), and all values are summarized in Methods, Fig. 7.25. In all the graphs the error bars or shaded areas are standard errors calculated from bootstrapping.

pared to that of untreated cells, Methods, Fig. 7.12), but RNAP was still significantly clustered (Fig. 7.15A) compared to free PAmCherry (Methods, Fig. 7.5). The number of RNAP clusters per cell decreased (1.9 clusters per cell, Fig. 7.15A, B, Methods, Fig. 7.22 and 7.23), their sizes reduced (104 nm, Fig. 7.15C, D), but they contained a similar fraction of total detected RNAP molecules compared to those in untreated cells (Fig. 7.15E, Methods, Figure 7.22 and 7.23). The averaged cellular localizations of all RNAP molecules in SHX-treated cells also exhibited a two-lobed distribution, although the middle cleft was less distinct compared to WT cells (Fig. 7.15F). Note that the nucleoid morphology and volume in SHX-treated cells was not significantly different from that of the WT cells (Methods, Fig. 7.5A, B, F). These results suggest that a high level of rRNA synthesis, as that in the rich medium growth condition, is not necessary for the formation of RNAP clusters.

7.5 RNAP forms clusters in the presence of only one rrn operon per chromosome

One possibility to explain the presence of a significant level of RNAP clusters in SHX-treated cells was that RNAP clusters might remain associated with multiple rrn operons that may spatially colocalize with each other [321], despite the lack of high transcription activity from these operons. To examine this possibility, we used a Δ6rrn strain, in which six out of seven rrn operons (except for rrnC) were removed from the chromosome [322]. This strain also contained an additional plasmid ptRNA67 [323] to provide tRNA genes in trans [324]. The Δ6rrn strain grew at a slower rate than WT cells under the

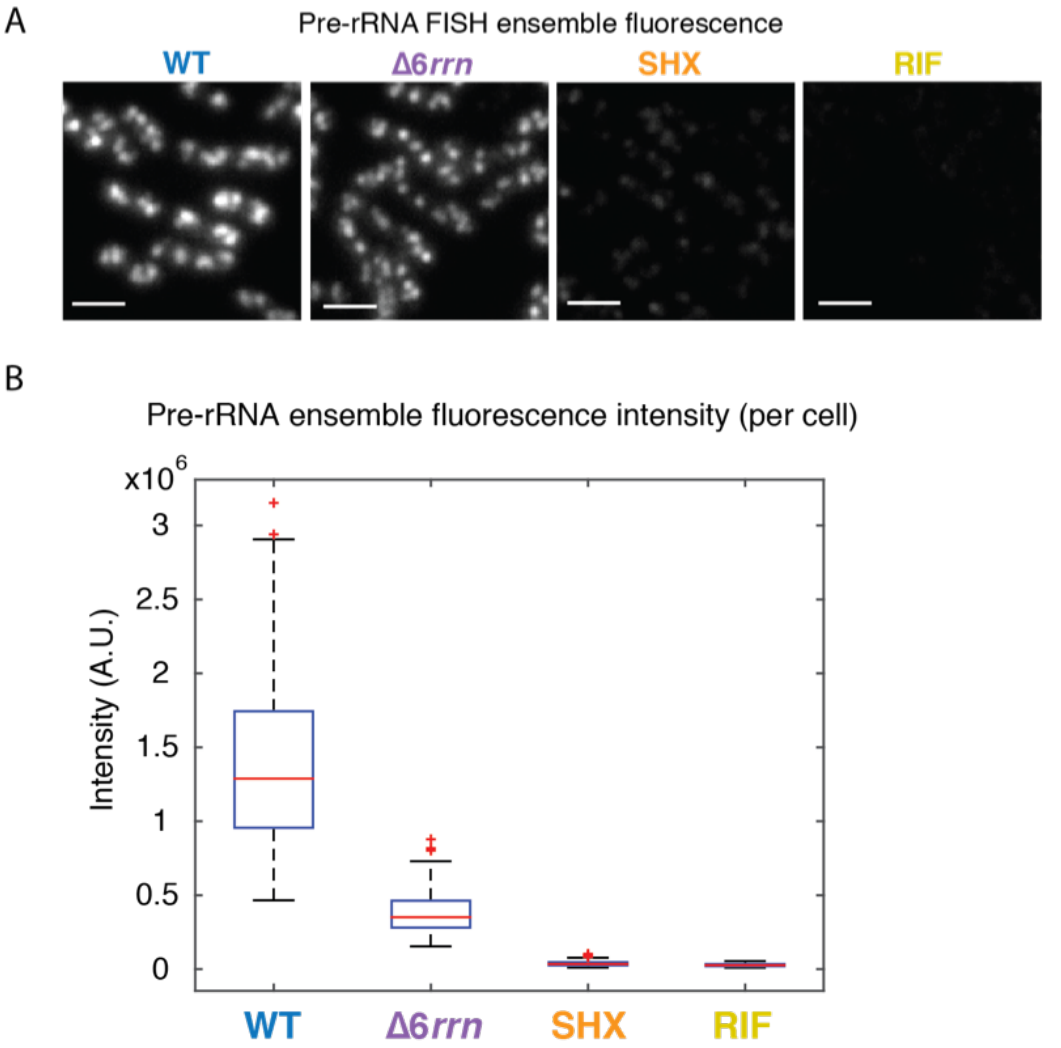

Figure 7.12: Pre-rRNA FISH signal under different conditions. (A) Ensemble pre-rRNA FISH fluorescence (large field view) of cells under different conditions. All the images are of the same contrast. Scale bar, 2 μm. (B) Integrated fluorescence intensity of pre-rRNA FISH signal of individual cells under different conditions are plotted as box plots, with the mean as the red line, and the boxed regions as the 25th and 75th percentiles, and outlier points are in red. WT: n = 72 cells, Δ6rrn: n = 72 cells, SHX: n = 110 cells, RIF: n = 76 cells.

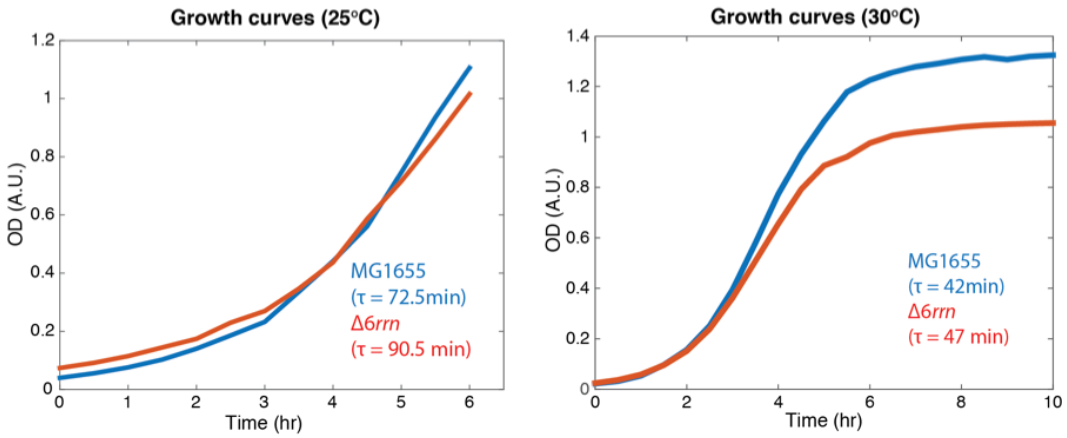

Figure 7.13: Comparison of the growth curve of Δ6rrn strain with WT strain MG1655 in EZ rich defined medium at both RT (25°C) and 30°C.

same rich medium growth condition (cell doubling time = 91± 1 min, Methods, Fig. 7.13), and showed a significant reduction in total rRNA synthesis (28% of WT cells, Methods, Fig. 7.12). However, the cellular distribution of RNAP and the properties of RNAP clusters in the Δ6rrn strain were remarkably similar to those of SHX-treated cells (Fig. 7.15G-L, Methods, Fig. 7.22, 7.23), and remained highly colocalized to residual pre-rRNA clusters (Methods, Fig. 7.14, 7.25). Additionally, we found that the nucleoid morphology and the total nucleoid volume of these cells were comparable to WT cells (Methods, Fig. 7.5C, F). These results suggest that the formation of RNAP clusters did not require a high level of rRNA synthesis activity or the presence of multiple rrn

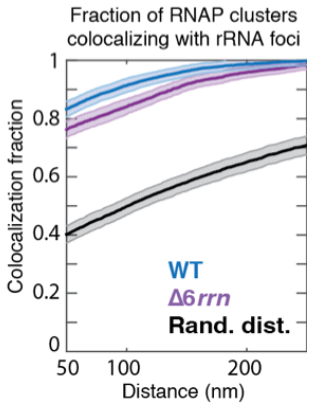

Figure 7.14: Fraction of RNAP clusters colocalizing with pre-rRNA clusters at different distances from 50 to 250 nm in the Δ6rrn strain. The black curve is the simulated basal level of colocalization due to chance. The blue curve is that of the WT under the rich medium growth condition plotted for comparison. The plotted colocalization fraction is corrected for detection efficiency of pre-rRNA clusters (Methods, Fig. 7.9A). The shaded areas are standard errors calculated from bootstrapping.

operon coding regions.

7.6 RNAP forms clusters in σ70-sequestered cells

Next, to probe the possibility that other non-rRNA transcription activities may contribute to the formation of RNAP clusters under the conditions of significantly reduced rRNA synthesis, we inhibited transcription from all housekeeping σ70 promoters by overexpressing a 10-kD bacteriophage T4 anti-σ protein AsiA [325] from an arabinose-inducible plasmid. AsiA binds to σ70 in the RNAP holoenzyme and prevents the holoenzyme from initiating transcription from σ70 promoters, which constitute about 75% of total *E. coli* promoters [326]. In these cells, we observed significantly elongated cells after

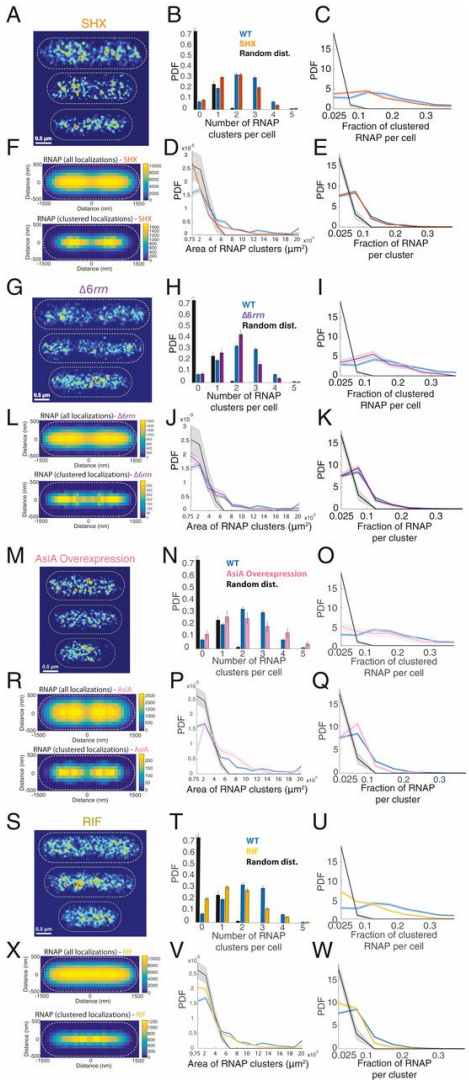

Figure 7.15: Characterization of RNAP clusters in live *E. coli* cells treated with SHX (A-F), in a rrn deletion strain (Δ6rrn, G-L), in cells with an overexpression of AsiA (M-R), and in cells treated with the global transcription inhibitor rifampicin (S-X) (See the following page for additional details).

Figure 7.15: Characterization of RNAP clusters in live *E. coli* cells treated with SHX (A-F), in a rrn deletion strain (Δ6rrn, G-L), in cells with an overexpression of AsiA (M-R), and in cells treated with the global transcription inhibitor rifampicin (S-X). (A, G, M, S) Representative superresolution images of RNAP-PAmCherry. Scale bar, 0.5μm. (B, H, N, T) Distribution of the number of RNAP clusters per cell, PDF is probability density function. (C, I, O, U) Distribution of the fraction of clustered RNAP per cell. (D, J, P, V) Distribution of the area of RNAP clusters. (E, K, Q, W) Distribution of the fraction of RNAP localizations per cluster. (F, L, R, X) 2D histogram of all RNAP localizations in a standard 3 μm x 1 μm cell (top), 2D histogram of only clustered RNAP localizations in a standard 3 μm x 1 μm cell (bottom). In (B-E, H-K, N-Q and T-W) the blue bars/curves are those of the WT under the rich medium growth condition for comparison, and the black curves are those calculated from simulated random distributions using the same number of localizations in the same cell volume for all the cells under each condition. All the mean values of these graphs are summarized in Fig. 7.22. In all the graphs (B-E, H-K, N-Q and T-W), the error bars or shaded areas are standard errors calculated from bootstrapping.

a two-hour arabinose induction (Methods, Fig. 7.16), indicating the detrimental effect of shutting down σ70 promoters [327]. However, while a smaller fraction of RNAP formed smaller clusters in AsiA-overexpressing cells compared to WT cells (Fig. 7.15M-R, 7.22, 7.23), these clusters were significant compared to that of free PAmCherry or random distribution (Methods, Fig. 7.5). The cellular distribution of all RNAP localizations appeared to expand and occupy a larger nucleoid volume compared to that in other conditions (Fig. 7.15R), likely indicating the presence of an increased fraction of free RNAP outside of the nucleoid. These results suggest that the formation of the

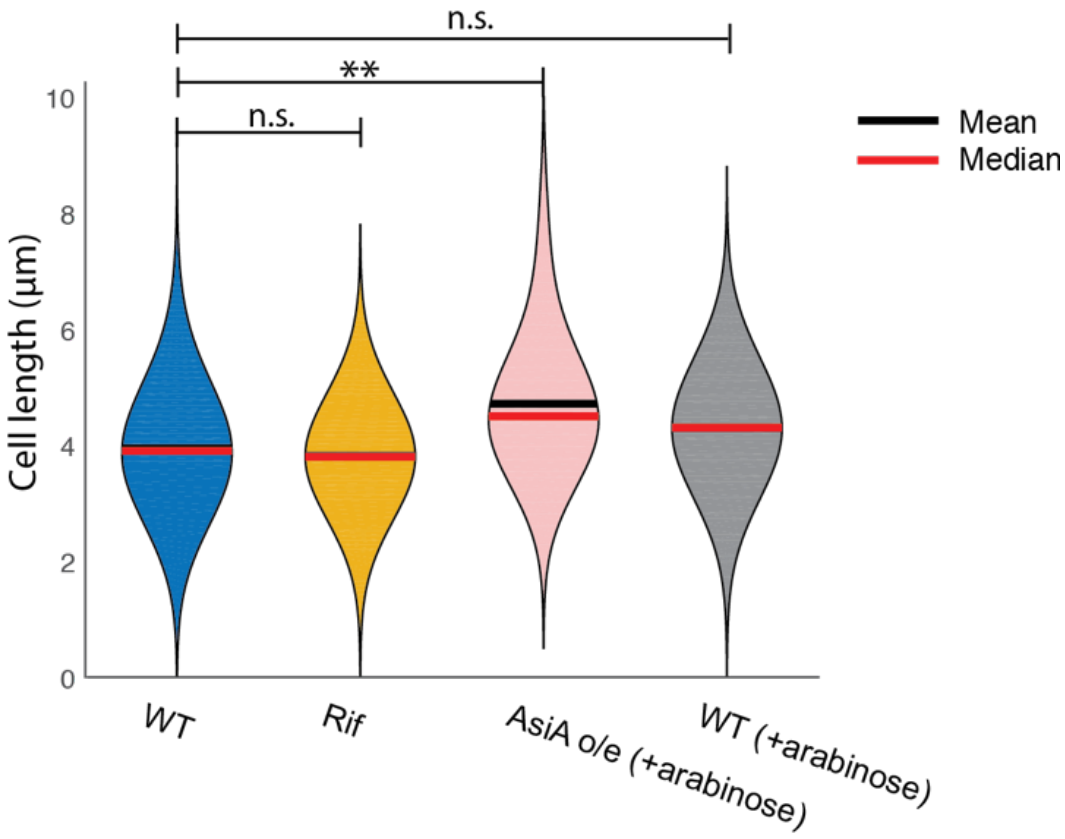

Figure 7.16: Comparison of cell lengths under the rich medium growth condition (EZRDM), in rifampicin-treated cells (Rif) and in AsiA-overexpressing cells. Cells induced with arabinose without the AsiA expression plasmid was used as a control. Asterisks indicate **: $p < 0.001$, n.s.: not significant.

remaining RNAP clusters did not require σ70 promoter activities.

7.7 RNAP clusters are significantly reduced in rifampicin-treated cells

In all the perturbation experiments described above, there still existed low levels of transcription activity (for example, mRNA transcription in SHX-treated cells and alternative σ factor transcription in AsiA overexpressing cells). Therefore, it is possible that RNAP might still form clusters engaged in remaining transcription. To test this possibility, we incubated cells with a global transcription inhibitor rifampicin (RIF, 100 μg ml^{-1}, 2 hours, Methods, Fig. 7.12). Rifampicin caps the RNA exit channel on RNAP before the nascent RNA chain emerges, and hence blocks transcription initiation and traps RNAP in an abortive cycle on the promoter [306]. We reasoned that if the majority of RNAP clusters we observed so far were indeed complexes involved in active transcription, these complexes would eventually finish transcription and run off when transcription initiation is inhibited, leading to the disappearance of RNAP clusters. Conversely, if RNAP clusters were long-lived complexes not involved in active transcription, they would persist despite the global inhibition of transcription. In rifampicin-treated cells, we observed that although the clustering of RNAP-PAmCherry was not completely eliminated compared to that of free PAmCherry control or the random distribution simulation, the extent of clustering was substantially reduced compared to all other conditions (Methods, Fig. 7.15S-X, 7.22, 7.23). The number and size of RNAP clusters decreased significantly (Fig. 7.15T, V and W), with more cells containing

fewer clustered RNAP molecules than that in WT cells (Fig. 7.15U). These results could suggest that most RNAP clusters were active transcription complexes under the rich medium growth condition, and that global transcription activity was a major contributor to the formation of RNAP clusters. However, we could not exclude the possibility that the observed changes were due to the expansion of the nucleoid under the condition of global transcription inhibition, upon which bound RNAP clusters dispersed [328]. Supporting this latter possibility was the observation that the cellular distribution of RNAP exhibited a homogenous, single-lobed pattern without discernible central cleft (Fig. 7.15X), mimicking that of the nucleoid under the same condition (Methods, Fig. 7.5D). Because both the global transcription activity and the nucleoid structure were significantly altered in cells treated with rifampicin, it is necessary to investigate RNAP's distribution under conditions where the effect of the nucleoid structure could be isolated without interfering with the transcription activity of rRNA.

7.8 Inhibition of gyrase activity leads to a redistribution of RNAP clusters and rRNA synthesis sites

As we described above, the cellular distribution of RNAP closely mimicked that of the corresponding nucleoid structure visualized using 3D SIM imaging under all the conditions tested (Methods, Fig. 7.5). These observations suggested that the spatial organization of these clusters might reflect that of the underlying nucleoid organization. To isolate the effect of the nucleoid structure

on RNAP distribution that is independent of transcription activity of rRNA, we decided to perturb the nucleoid organization by inhibiting gyrase activity.

The *E. coli* chromosome is highly compact and organized at different levels from topological domains to macrodomains (MDs) [329, 330, 331]. These organizations likely dictate the spatial arrangement of different DNA segments, upon which RNAP may preferentially bind and form clusters. Negative supercoiling is a major chromosome compacting factor, and it is only introduced by gyrase, a type II topoisomerase in *E. coli* [332]. We thereby examined specifically the effect of DNA supercoiling on the spatial organization of RNAP.

We treated WT cells with a gyrase inhibitor novobiocin (NOV, 300 μg ml^{-1} for 30 min) and performed two-color superresolution imaging using pre-rRNA FISH and RNAP-PAmCherry (Fig. 7.17A-D). Novobiocin inhibits gyrase activity by abolishing ATP binding to the ATPase domain in the GyrB subunit [333, 334]. We found that the average rRNA synthesis activity per cell was not significantly affected by the inhibitor, as the total intensity of pre-rRNA FISH signal remained similar to untreated cells (Methods, Fig. 7.18A), even when high inhibitor concentrations and long-time treatment were used (Methods, Fig. 7.18B). We further verified that the persistent rRNA synthesis during gyrase inhibition was not due to altered rRNA degradation kinetics in the presence of the inhibitor (Methods, Fig. 7.18C). The minimal effect of gyrase inhibition on rRNA synthesis has been observed previously [335, 336], although conflicting results have been reported as well [337, 338]. Interestingly, while the total pre-rRNA FISH signals remained unchanged under our experimental condition, we observed a greater number (on average 5.5 per cell compared to

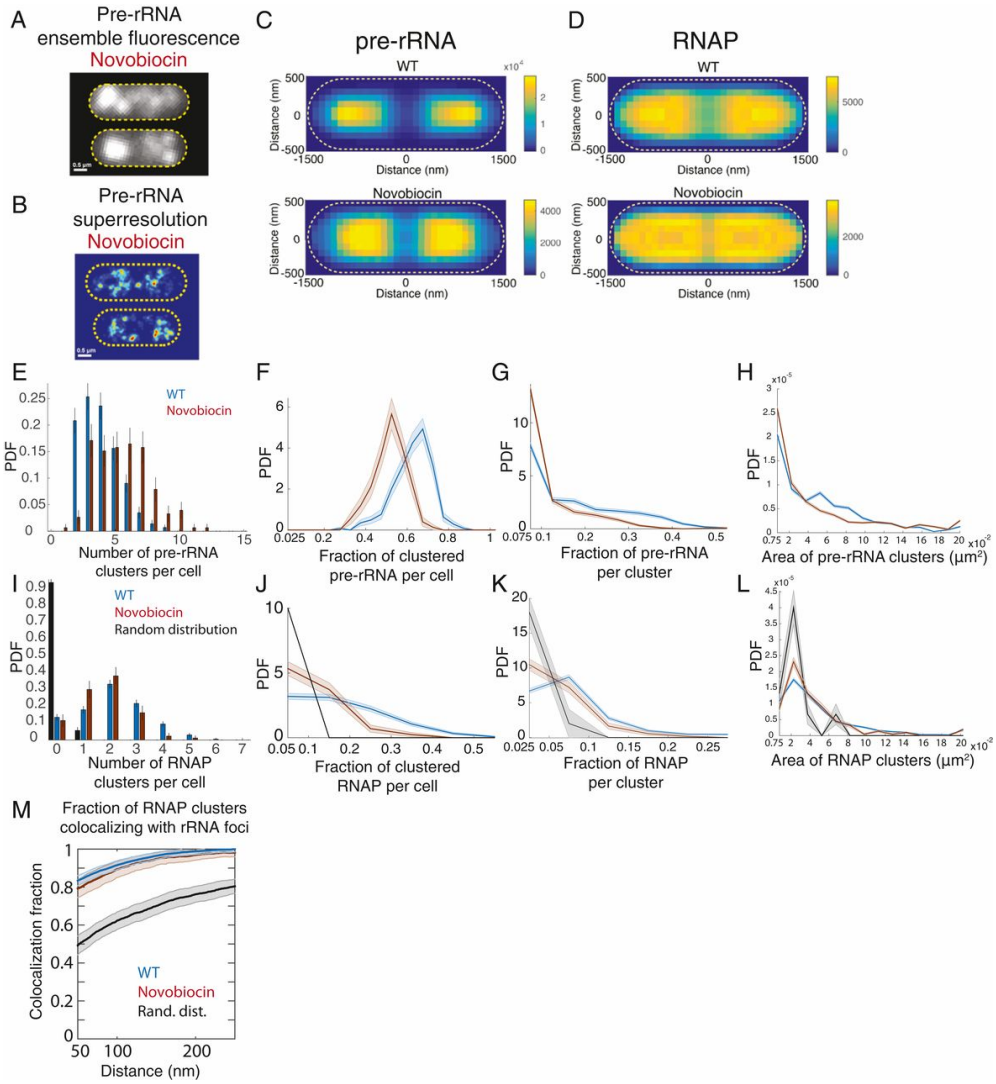

Figure 7.17: Inhibition of gyrase activity led to dispersed distributions of RNAP and pre-rRNA. (See the following page for additional details.)

Figure 7.17: Inhibition of gyrase activity led to dispersed distributions of RNAP and pre-rRNA. (A) Ensemble fluorescence of Pre-rRNA FISH signal in fixed, novobiocin-treated cells. Individual cells are outlined in yellow. (B) Representative superresolution images of pre-rRNA distribution in fixed, novobiocin treated cells. (C) 2D histograms of all pre-rRNA localizations in a standard 3 μm x 1 μm fixed cell under the rich medium growth condition (top) and in cells treated with novobiocin (bottom). (D) 2D histograms of all RNAP localizations in a standard 3 μm x 1 μm fixed cell under the rich medium growth condition (top) and in cells treated with novobiocin (bottom). (E-L) Distributions of properties of pre-rRNA (E-H) and RNAP clusters (I-L) in novobiocin-treated cells, PDF is probability density function. (E, I): Distribution of the number of clusters per cell. (F, J): Distribution of fraction of clustered pre-rRNA (F) or RNAP (J) per cell. (G, K): Distribution of fraction of pre-rRNA (G) or RNAP (K) localizations per clusters. (H, L): areas of clusters. (M) Fraction of RNAP clusters colocalizing with pre-rRNA clusters in novobiocin-treated cells. In all plots the WT rich medium growth conditions are plotted in blue for comparison; novobiocin-treated conditions are in dark red, and the background colocalization levels using simulated images are in black. All error bars or shaded areas are standard error calculated using bootstrapping. All the mean values of these graphs are summarized in Methods, Fig. 7.24 and 7.25.

3.9 in untreated cells) of less dense (52% of total cellular localizations compared to 63% in untreated cells) pre-rRNA clusters (Fig. 7.17E-H, Methods, Fig. 7.24, 7.23). RNAP clusters persisted in these gyrase-inhibited cells as well (Fig. 7.17D), remained highly colocalized with pre-rRNA clusters (Fig. 7.17M), but contained fewer RNAP molecules (Fig. 7.17I to L, Methods, Fig. 7.23). Interestingly, the cellular distributions of pre-rRNA and RNAP clusters in gyrase inhibited cells exhibited a similarly, spatially dispersed pattern compared to that of untreated cells (Fig. 7.17C, D), and the average positioning of RNAP clusters and pre-rRNA clusters moved 80 nm radially toward the nucleoid periphery (Methods, Fig. 7.19). Note that the cellular distribution RNAP again mimicked that of the expanded nucleoid under the same condition (Methods, Fig. 7.5E, F). A different gyrase inhibitor, nalidixic acid (NA, 50 μg ml^{-1} for 10 min), which acts on the GyrA subunit by stabilizing the DNA-cleaved complex, produced a similar effect on the cellular distributions of pre-rRNAs and RNAP (Methods, Fig. 7.20A to J and O) but less on the properties of RNAP clusters (Methods, Fig. 7.20 K to N), likely due to the short time (10 min) used to treat cells in order to avoid double stranded ds-DNA breaks. The significant redistribution of RNAP and pre-rRNA clusters in the presence of altered nucleoid structure but unchanged rRNA transcription activity suggested that the characteristics and organization of the nucleoid, here in particular compaction by negative supercoiling, play a large role in the

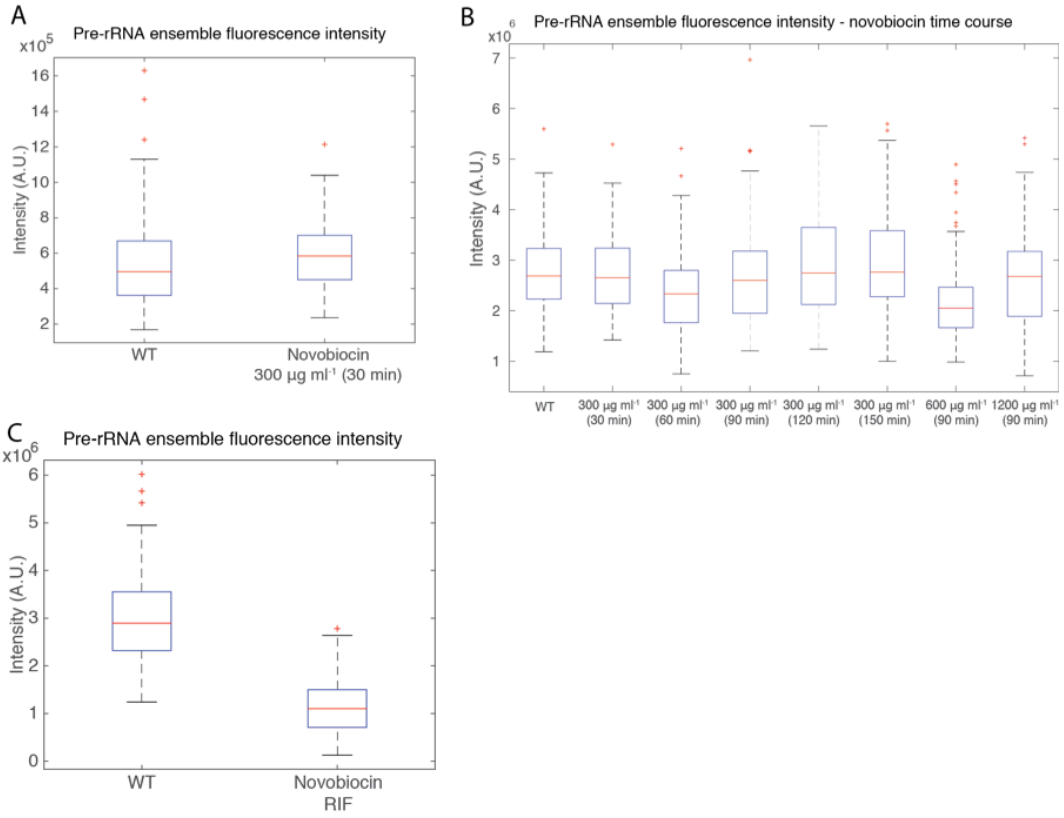

Figure 7.18: Pre-rRNA ensemble fluorescence intensities under gyrase inhibited conditions. (A) Quantification of cellular pre-rRNA signal intensities for WT (n = 134 cells) and novobiocin (30 min 300 μg ml^{-1}, n = 105 cells) treated cells. (B) A time series of novobiocin treatment (0-150 min, 300 μg ml^{-1}), with higher concentration of novobiocin also used (600 μg ml^{-1} and 1200 μg ml^{-1}, 90 min); WT: n = 84 cells; 300 μg ml^{-1}, 30 min: n = 80 cells; 300 μg ml^{-1}, 60 min: n = 65 cells; 300 μg ml^{-1}, 90 min: n = 80 cells; 300 μg ml^{-1}, 120 min: n = 74 cells; 300 μg ml^{-1}, 150 min: n = 72 cells; 600 μg ml^{-1}, 90 min: n = 96 cells; 1200 μg ml^{-1}, 90 min: n = 81 cells. (C) Quantification of cellular pre-rRNA signal intensities for WT (n = 58 cells), and for cells followed by additional 10-min RIF treatment without washing out novobiocin (100 μg ml^{-1}, n = 66 cells). All means are shown as red lines, the boxed regions at the 25th and 75th percentiles, and outliers' points are in red.

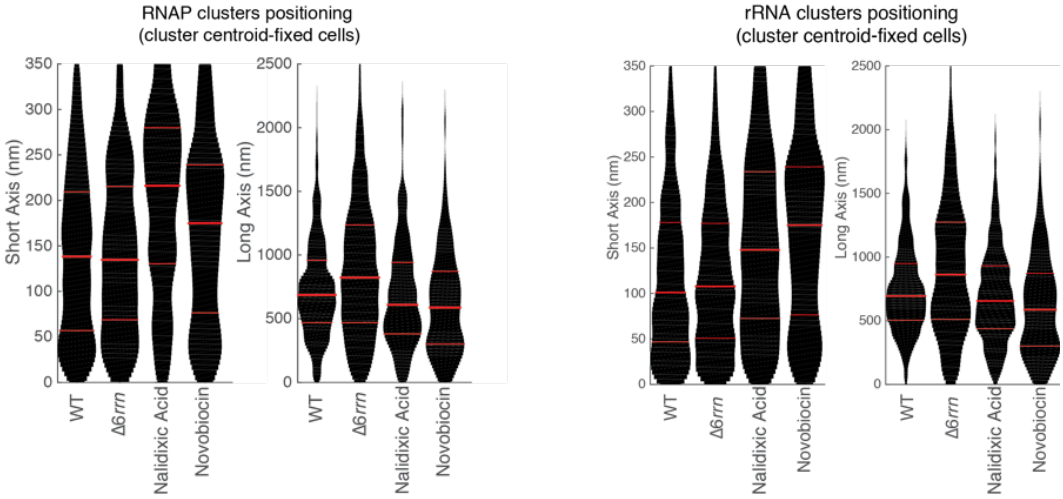

Figure 7.19: Averaged cellular positioning of the centroids of RNAP clusters (left) or centroids of rRNA clusters (right) projected along the long and short axes of cells. All data are from fixed cell experiments and all cell sizes are normalized to a standard cell size of 1 μm x 3 μm. Cell center is defined as (0,0). Means are shown as middle red lines in the distributions, with 25th and 75th percentiles shown as flanking red lines. In the main text, these distances were converted back to 3D radial distances by dividing a projection factor 0.64 [169].

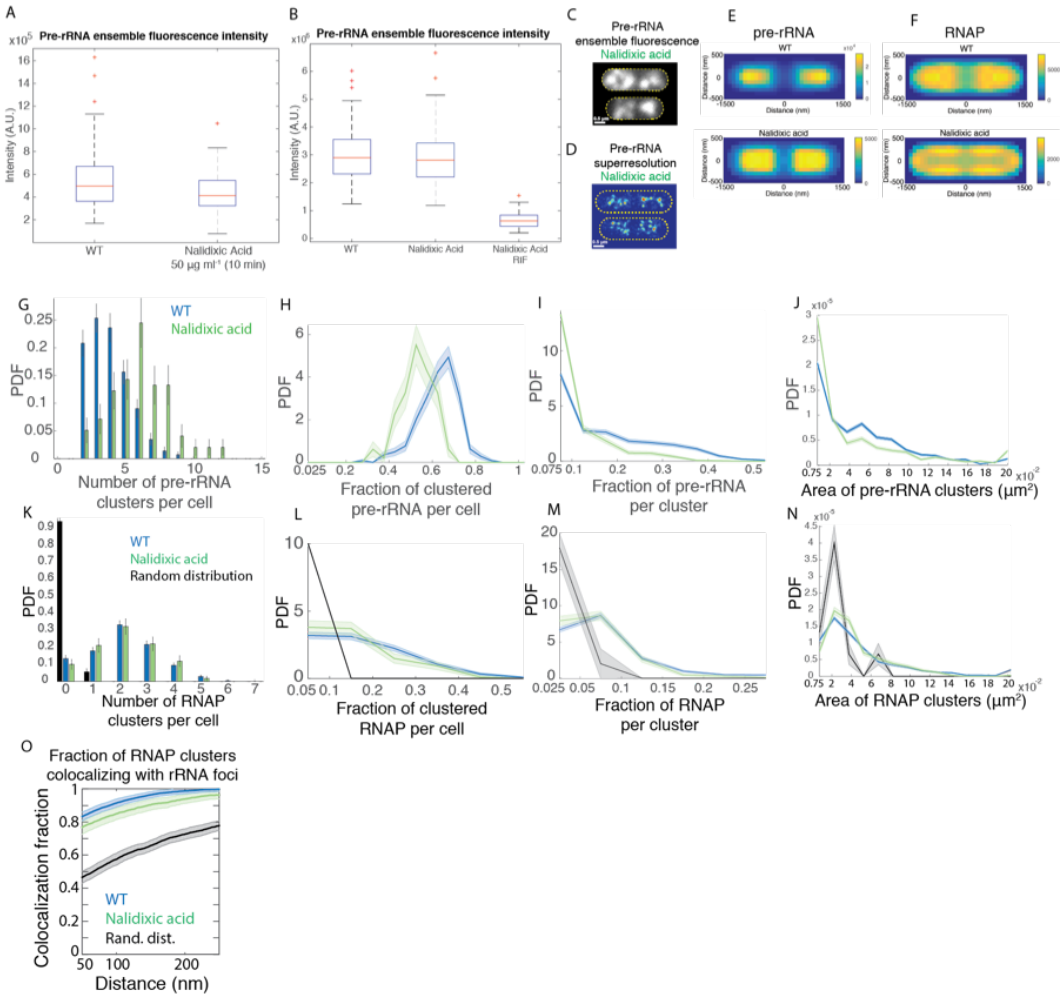

Figure 7.20: RNAP and pre-rRNA characterizations in nalidixic acid treated cells. (See the follow page for details)

Figure 7.20: RNAP and pre-rRNA characterizations in nalidixic acid treated cells. (A) Quantification of cellular pre-rRNA signal intensities for WT (n = 134 cells) and nalidixic acid (10 min 50 μg ml^{-1}, n = 102 cells) treated cells. (B) Quantification of cellular pre-rRNA signal intensities for WT (n = 58 cells), nalidixic acid (10 min 50 μg ml^{-1}, n = 61 cells), and an additional condition with a 10 min RIF treatment (100 μg ml^{-1}) follow-up without washing out the gyrase inhibitor (n = 57 cells). All means are shown as red line, with the boxed region at the 25th and 75th percentiles, and outlier points are in red. (C) Ensemble fluorescence of Pre-rRNA FISH signal in nalidixic acid-treated cells. Individual cells are outlined in yellow. Scale bar, 0.5 μm. (D) Representative superresolution images of pre-rRNA distribution in nalidixic acid treated cells. Scale bar, 0.5 μm. (E) 2D histograms of all pre-rRNA localizations in a standard 3 μm x 1 μm fixed cell under the rich medium growth condition (top) and in cells treated with and nalidixic acid (bottom). (F) 2D histograms of all RNAP localizations in a standard 3 μm x 1 μm fixed cell under the rich medium growth condition (top) and in cells treated with nalidixic acid (bottom). (G-N) Distributions of properties of pre-rRNA (G-J) and RNAP clusters (K-N) in gyrase-inhibited cells. (G, K): Distribution of the number of clusters per cell, PDF is probability density function. (H, L): Distribution of fraction of clustered pre-rRNA (H) or RNAP (L) per cell. (I, M): Distribution of fraction of pre-rRNA (I) or RNAP (M) localizations per cluster. (J, N): areas of clusters. (O): Fraction of RNAP clusters colocalizing with pre-rRNA clusters in nalidixic acid-treated cells. In all plots, the WT rich medium growth conditions are plotted in blue for comparison nalidixic acid-treated cells (in green), and the background colocalization levels using simulated images are in black. All shaded areas are standard error calculated using bootstrapping. All the mean values of these graphs and their statistical significance from untreated cells are summarized in Methods, Fig. 7.23, 7.24 and 7.25.

spatial distribution of RNAP clusters.

7.9 Discussion

In this study, we investigated the prokaryotic transcription factory model in detail using a combination of quantitative superresolution imaging and perturbation analyses. Below we compare our results with previous work and discuss the implications of this work.

7.9.1 Spatial organization of RNAP

We observed that in *E. coli* cells grown in rich medium, RNAP was spatially organized into large, dense clusters occupying the same area as the nucleoid. These clusters had a radius of 130 nm (Fig. 7.1H), and could be made up of collections of multiple small RNAP clusters observed in a previous study [40]. Given a total cellular level of RNAP at 5000 molecules per cell [339, 314, 315] under a similar growth condition, and that majority (90%) of cellular RNAP remain bound on DNA [51, 236], we estimated that each RNAP cluster contained 350 molecules. The cellular distribution of all RNAP molecules exhibited a two-lobed pattern with a clear cleft in the middle (Fig. 7.1B), mimicking that of two replicated and segregated nucleoids (Methods, Fig. 7.5A). Compared to the distribution of all RNAP molecules, RNAP clusters were tighter and more concentrated toward the center of the two lobes with an average distance of 120 nm radially from the center of the cell (Methods, Fig. 7.21). This observation is consistent with previous superresolution studies of the spatially separated ribosomes and RNAP in *E. coli* [89] —ribosome at

the nucleoid boundary while RNAP predominately at the center. We have shown that with respect to HU clusters under the same growth condition, RNAP clusters were positioned more peripherial within the nucleoid territory (Methods, Fig. 7.7), in agreement with a previous study [40]. In addition, under faster growth conditions (37°C LB and 37°C EZRDM), RNAP clusters also shifted their distribution even more toward the center of the cell along the short axis (Methods, Fig. 7.4D), indicating that their distribution was sensitive to growth rate. In previous studies, the apparent spatial segregation between ribosome and RNAP has been used to question the coupling between transcription and translation in bacterial cells [89, 51]. It is possible that periphery-localized small RNAP clusters, which may be undetectable in our stringent distance-based clustering algorithm, contained RNAP molecules actively engaged in mRNA transcription that is coupled to translation, while the larger, nucleoid center-localized RNAP clusters we observed were responsible for rRNA synthesis, which does not require translation.

7.9.2 Spatial organization of pre-rRNA clusters

We used a pre-RNA FISH probe targeting the leader sequence of the 16S rRNA to detect rRNA transcription activity. Because newly synthesized pre-rRNAs are processed before they are incorporated into ribosomes, the pre-rRNA probe marks new rRNA synthesis sites. Compared to RNAP, pre-rRNAs formed similarly sized (130 nm in radius) but denser (containing $> 60\%$ of detected cellular pre-rRNAs) clusters. The overall cellular distribution of pre-rRNAs

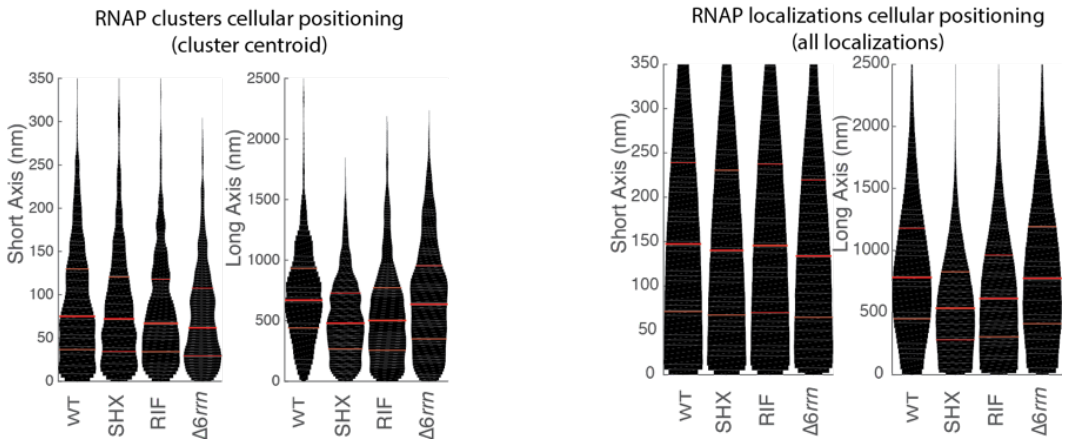

Figure 7.21: Averaged cellular positioning of the centroids of RNAP clusters (left) or all RNAP localizations (right) along the long and short axes of cells. All data are from live cell experiments and all cells' sizes are normalized to a standard cell size of 1 μm x 3 μm. Cell center is defined as (0,0). Means are shown as middle red lines in the distributions, with 25th and 75th percentiles shown as flanking red lines. In the main text, these distances were converted back to 3D radial distances by dividing a projection factor 0.64 [169].

also exhibited a two-lobed pattern, but relatively more concentrated at the center of the nucleoid compared to RNAP clusters in fixed cells (Methods, Fig. 7.19). On average we observed four pre-rRNA clusters per cell or two per chromosome (Fig. 7.11C). Because four out of the seven rrn operons reside close to the replication origin oriC on the chromosome, it is possible that the two pre-rRNA clusters reflected two copies of replicated oriC for each chromosome, and thus most cells contained two copies of the chromosome with four copies of oriC region, consistent with previous observations when the Ori region was labeled [340]. Alternatively, it is also possible that the two pre-rRNA clusters reflected two groups of transcribing rrn operons on the same copy of the chromosome that are spatially distinguishable from each other under our resolutions. A recent study found that, except for rrnC, all the rrn operons are within a spatial distance of 80 to 130 nm (median of distributions) to each other [321], but it remains unknown whether these rrn operons indeed co-occupy the same area in the nucleoid and whether they could be accommodated in one pre-rRNA cluster (radius of 130 nm). Because of the nearly identical pre-rRNA sequences of all the rrn operons, we could not design a probe with high confidence to distinguish the transcription activity of individual rrn operons, and hence further investigation is required to address whether each pre-rRNA cluster reflects the transcription activity from an individual or a collection of rrn operons.

7.9.3 The contribution of transcription activity to the spatial organization of RNAP

We observed that under the rich medium growth condition, RNAP clusters highly colocalized with pre-rRNA FISH probe signals. This suggests that under the rich medium growth condition the majority of RNAP clusters were actively transcribing rRNAs. These results are the most direct demonstration of the activity of RNAP clusters for the rich medium growth condition. Previous studies have assumed but not validated that RNAP clusters are transcribing rRNAs in fast-growing cells [137, 292]. Surprisingly, when we abolished rRNA transcription using serine hydroxamate, reduced rRNA transcription to 1/3 of WT cells using a mutant strain lacking six of the seven rrn operons (Δ6rrn), or inhibited all σ70-promoter activities by overexpressing AsiA, we found that there were still significant levels of RNAP clustering (Fig. 7.15A-R). The drastically different transcription activities but similar organizations of RNAP under the three different conditions hence suggest that rRNA transcription activity may not be the main driving force for the organization of RNAP clusters under our experimental conditions. Furthermore, because there was only one copy of the rrnC operon in the Δ6rrn strain, it is unlikely that multiple rrn operons were required for the formation of RNAP clusters as previously proposed [289]. The greatest perturbation to RNAP clusters was seen under the condition where global transcription initiation was inhibited by rifampicin (Fig. 7.15S-X), although under this condition the nucleoid structure was also altered, and RNAP still exhibited a clustered distribution different from the random distribution. Note that our results did not automatically im-

ply that these persisting RNAP clusters bind to the same chromosomal DNA sequences, have the same molecular compositions, or localize to the same cellular positions. Further biochemical investigations of these properties will be needed.

7.9.4 The contribution of nucleoid structure to the spatial organization of RNAP

In all of our experiments we observed that the cellular distribution of RNAP mimicked the shape of the underlying nucleoid irrespective of rRNA synthesis activity (Methods, Fig. 7.4). We thereby investigated the contribution of the nucleoid structure on the spatial distribution of RNAP by inhibiting gyrase. Gyrase is the only type II topoisomerase in *E. coli* that introduces negative supercoiling into the chromosome, which is the major force in compacting the nucleoid [332, 341]. In gyrase-inhibited cells via both nalidixic acid and novobiocin, we observed similar levels of pre-rRNA signal (Methods, Fig. 7.18A) but saw a significant shift in the cellular positioning of RNAP and pre-rRNA clusters, both of which expanded 120 nm radially toward the nucleoid periphery (Fig. 7.17C, D, Methods, Fig. 7.19, 7.20E, F). Previous studies have documented that gyrase inhibition affects the expression of more than three hundred mRNA genes that are sensitive to supercoiling [342, 343], but produced mixed results on the effect of rRNA transcription [338, 337, 336]. The pre-rRNA FISH probe detects the 5' leader sequence of 16S rRNA, hence the unchanged FISH signal in gyrase-inhibited cells only reflected unaltered rRNA transcription initiation. It is possible that rRNA elongation was inhibited due

to accumulated positive supercoiling ahead of transcription in the absence of gyrase. In such a case, we should expect that after a long inhibition time, the rRNA transcription initiation rate would gradually decrease as accumulated positive supercoiling eventually inhibits transcription initiation, which was demonstrated previously for the production of mRNA in vitro [82]. However, we observed similar pre-rRNA signal even after we incubated cells with high concentrations of novobiocin (300 to 1200 μg ml^{-1}) for extended periods (90 to 150 min, Methods, Fig. 7.18B), demonstrating that the total rRNA transcription activity in these cells was minimally affected. Therefore, these experiments most likely suggested that the nucleoid structure contributes significantly to the spatial organization of RNAP. It is possible that a relaxed chromosome repositioned different DNA segments (upon which RNAP clusters form) to occupy a larger cell volume, as suggested by SIM imaging of the gyrase-inhibited nucleoid (Methods Fig. 7.4E, F). Clearly, further studies, such as genetic and biochemical perturbations of nucleoid-organization factors, are required to investigate the effect of nucleoid structure on the spatial organization of RNAP.

In summary, our study demonstrated that there was a rRNA transcription activity-independent spatial organization of RNAP in *E. coli* and that the underlying nucleoid structure played important roles in organizing RNAP clusters. Further experiments investigating the molecular nature and function of RNAP clusters are required. In particular, we do not know whether these RNAP clusters we observed are associated with specific chromosomal DNA sequences, or whether they are self-promoting oligomeric complexes similar to

liquid droplets observed in eukaryotic cells, which are mediated by multivalent interactions among proteins and nucleic acids [344, 345]. Intriguingly, it has been shown that a small regulatory ncRNA 6S can sequester σ70 holoenzyme; these RNAP-RNA complexes may also contribute to the clusters we observe under conditions where transcription activity from σ70 promoters is low, although currently there has been no report of a clustered 6S RNA distribution [346, 347]. Furthermore, we do not know the biological significance of RNAP clusters. In eukaryotic cells, it was suggested that RNAP clusters might represent pre-formed transcription complexes that are "poised" ready for rapid transcription induction [296, 297, 298, 299]. In bacterial cells, such a role has not been demonstrated, but studies have shown that there are typically higher levels of RNAP association at promoter and promoter-like sequences compared to within coding sequences [300, 301, 302, 303, 304, 305]. Perhaps looking at the colocalization of RNAP with important transcription regulators (NusA [348, 349], NusB [350, 351], NusE [352], NusG [300, 353], and SuhB [354], *etc.*) that interact with RNAP under different conditions would help elucidate the molecular makeup and functional significance of RNAP clusters. Regardless, further investigations into the spatial organization of RNAP in small bacterial cells will certainly bring in new knowledge complementing in-vitro biochemical and in-vivo genetic studies of prokaryotic transcription.

RNAP cluster characteristics in live cell superresolution images under various conditions.
Values shown are mean ± standard error, with (N) as the number of data points used in measurements.

Condition (live cell)	# of RNAP clusters per cell	Fraction per cluster	Fraction in clusters per cell	Cluster area (μm^2) – cluster radius (nm)
Neg. simulation	0.27 ± 0.04 (141)	0.044 ± 0.001 (38)	0.01 ± 0.002 (141)	0.020 ± 0.002 (38) – 83 ± 26 (38)
WT	2.13 ± 0.05 (664)	0.076 ± 0.001 (1385)	0.16 ± 0.005 (664)	0.052 ± 0.002 (1385) – 129 ± 25 (1385)
SHX	1.85 ± 0.04 (714)	0.075 ± 0.001 (1317)	0.14 ± 0.004 (714)	0.034 ± 0.001 (1317) – 104 ± 19 (1317)
RIF	1.54 ± 0.05 (559)	0.061 ± 0.001 (860)	0.09 ± 0.003 (559)	0.037 ± 0.001 (860) – 108 ± 21 (860)
AsiA overexpression	2.07 ± 0.16 (75)	0.065 ± 0.003 (155)	0.13 ± 0.003 (75)	0.056 ± 0.004 (155) – 134 ± 36 (155)
Δ6rrn	1.83 ± 0.08 (151)	0.070 ± 0.002 (277)	0.13 ± 0.006 (151)	0.040 ± 0.003 (277) – 113 ± 30 (277)
EZRDM (mEos3.2)	2.51 ± 0.07 (258)	0.086 ± 0.002 (648)	0.22 ± 0.007 (258)	0.077 ± 0.003 (648) – 157 ± 33 (648)

Figure 7.22: RNAP cluster characteristics in live cell superresolution images under various conditions.

7.10 Methods

7.10.1 Bacterial strains and constructions

The wild-type (WT) strain background was MG1655. The RpoC-PAmcherry (CC253) and RpoC-mEos3.2 (XW023) chromosomal fusion strains were constructed using Λ-RED-mediated homologous recombination [355]. Specifically, the linear fragment containing the fluorescent protein 'FP-frt-kanR-frt' sequence was first generated and subcloned into the pKD13 plasmid [355]. The 50-bp homologous flanking regions were then added to the linear fragment using primer pairs 15-16 and 17-18. The linear fragment was transformed into MG1655 cells containing the pKD46 plasmid with 0.2% L-arabinose induction. Recombinants grown on LB + kanamycin plates were verified by colony PCR. The pKD46 plasmid was next cured by growing cells at the restrictive temperature 37°C. The RpoC-PAmCherry fusion in the Δ6rrn strain (CC302)

was constructed similar to described above [307]. In later constructions, the frt-kanR-frt cassette was flipped out using the PCP20 plasmid [355] to generate strain ACL002. A chromosomal DNA site marker (tetO6) was inserted at different chromosomal locations of ACL002 to generate a series of dual-labeled strains (ACL066, ACL020, XW030, XW033, ACL036, XW017, and XW018) using primer pairs 1 to 14, and Λ-RED mediated homologous recombination as described above. A plasmid expressing the TetR-EYFP reporter was constructed from pZH102R33Y29 [356] and transformed into all the dual-labeled strains. We imaged both RNAP and DNA localizations of all the strains in live cells, but only included RNAP localization data in this work due to the limit of space. DNA localization data will be described in an accompanying study.

7.11 Methods

7.11.1 Cell growth

Single *E. coli* colonies were picked from freshly streaked LB plates and cultured in EZ Rich Defined Media (EZRDM, Teknova) with 0.4% glucose, at room temperature (RT) overnight with shaking. Antibiotics (kanamycin (Sigma-Aldrich 1355006) and carbenicillin (Sigma-Aldrich C3416)) were added at 50 μg ml^{-1} when appropriate. The next morning, cells were reinoculated (1:200) into fresh EZRDM with 0.4% glucose and grown at RT until they reached mid-log phase (O.D.600 0.3-0.4). To induce TetR-EYFP expression, cells were harvested and resuspended in fresh EZRDM supplemented with 0.3% L-arabinose and 0.4%

Tabulated *p*-values from two-sample *t*-tests and KS (Kolmogorov-Smirnov) tests for all reported RNA, pre-rRNA, HU, free PAmCherry cluster characteristics. All reported p-values are with respect to the WT (RT, EZRDM) condition. The two-sample *t*-test was used for the discreet distribution of the number of RNAP clusters per cell since KS test can only be applied to continuous distributions. (*: $p < 0.01$, **: $p < 0.001$, n.s.: not significant)

	Two-sample t-test	KS-test	KS-test	KS-test	Two-sample t-test	KS-test	KS-test	KS-test	KS-test	KS-test
Conditions	Number of RNAP clusters per cell	Frac. of clustered RNAP per cell	Frac. of RNAP per cluster	Area of RNAP clusters	Number of pre-rRNA clusters per cell	Frac. of clustered pre-rRNA per cell	Frac. of pre-rRNA per cluster	Area of pre-rRNA clusters	Short axis position (RNAP clusters)	Short axis position (pre-rRNA clusters)
Live cell conditions (compared to WT (RT, EZRDM))										
SHX (RT)	**	*	n.s.	**	-	-	-	-	n.s.	-
RIF (RT)	**	*	**	**	-	-	-	-	*	-
Δ6*rrn* (RT)	*	*	n.s.	n.s.	-	-	-	-	*	-
AsiA o/e (RT)	n.s.	*	n.s.	**	-	-	-	-	n.s.	-
Neg. simulation	**	*	**	**	-	-	-	-	**	-
Fixed cell conditions (compared to WT (RT, EZRDM))										
EZRDM (37 °C)	**	**	**	**	-	-	-	-	**	-
LB (37 °C)	**	**	n.s.	**	-	-	-	-	**	-
Δ6*rrn* (RT)	n.s.	**	**	*	**	**	**	**	n.s.	n.s.
Nalidixic acid (RT)	n.s.	n.s.	n.s.	n.s.	**	**	**	**	**	**
Novobiocin (RT)	n.s.	**	**	n.s.	**	**	**	**	*	**
Neg. simulation	**	**	**	n.s.	**	-	-	-	**	-
HU (RT)	**	**	**	n.s.					**	
Free PAmCherry (RT)	**	**	n.s.	n.s.					*	

Figure 7.23: Tabulated p-values from two-sample t-tests and KS (Kolmogorov-Smirnov) tests for all reported RNA, pre-rRNA, HU, free PAmCherry cluster characteristics.

glycerol and allowed to grow for two additional hours (this condition was used for all live cell imaging experiments reported in this work). For live cell experiments with drug-treatment, 2hr RIF inhibition (100 μg ml^{-1}) was performed after the 2hr TetR-EYFP induction, and 1hr SHX (500 μg ml^{-1}) treatment was performed during the last hour of TetR-EYFP induction. For AsiA overexpression experiments, the AsiA gene was under a PBAD promoter on a plasmid to minimize leaky expression. To induce AsiA expression, cells were harvested and resuspended in EZRDM supplemented with 0.4% L-arabinose and 0.4% glycerol, the cells were allowed to grow for two additional hours at RT. Live cells after induction or drug treatments were harvested and prepared for imaging as described in the section below. For fixed cell experiments, growth and drug treatments were done as follows: cells were grown to mid-log phase in EZRDM with 0.4% glucose at RT. Cells were treated with drugs when appropriate; SHX treatment was performed at 500 μg ml^{-1} for 1hr, RIF treatment was performed at 100 μg ml^{-1} for 2hr, and novobiocin treatment was performed with 300 μg ml^{-1} for 30 min. For faster growth conditions, cells were picked from freshly streaked LB plates and cultured in either LB media or EZ Rich Defined Media (EZRDM, Teknova) with 0.4% glucose, at 37°C overnight with shaking. The next morning cells were reinoculated with 1:200 dilution in the same fresh medium and allowed to grow at 37°C until the culture O.D.600 reached 0.3-0.4. Cells were then harvested and fixed in 3.7% (v/v) paraformaldehyde (16% Paraformaldehyde, EM Grade, EMS) for 15 min at RT, washed with 1X PBS and imaged immediately.

Pre-rRNA cluster and RNAP cluster characteristics in pre-rRNA-RNAP two-color superresolution imaging experiments (fixed cell). Values shown are mean ± standard error, with (N) as the number of data points used in measurements.

Condition	# of RNAP clusters per cell	Fraction per cluster (RNAP)	Fraction in clusters per cell (RNAP)	Cluster area (μm^2) – radius (nm) (RNAP)	# of Pre-rRNA clusters	Fraction per cluster (Pre-rRNA)	Fraction in clusters per cell (Pre-rRNA)	Cluster area (μm^2) – radius (nm) (Pre-rRNA)
Neg. simulation	0.06 ± 0.02 (176)	0.05 ± 0.004 (10)	0.003 ± 0.001 (176)	0.028 ± 0.005 (10) – 94 ± 41 (10)	-	-	-	-
WT	2.07 ± 0.06 (398)	0.08 ± 0.002 (823)	0.17 ± 0.006 (398)	0.050 ± 0.002 (823) – 125 ± 23 (823)	3.9 ± 0.q (288)	0.16 ± 0.004 (1086)	0.63 ± 0.005 (288)	0.051 ± 0.002 (1086) – 128 ± 22 (1086)
Δ6rrn	1.96 ± 0.07 (286)	0.07 ± 0.003 (560)	0.13 ± 0.005 (286)	0.057 ± 0.002 (560) – 135 ± 28 (560)	5.0 ± 0.1 (192)	0.14 ± 0.003 (939)	0.70 ± 0.006 (192)	0.053 ± 0.002 (939) – 131 ± 24 (939)
Nalidixic acid	2.10 ± 0.12 (108)	0.07 ± 0.003 (227)	0.15 ± 0.01(108)	0.043 ± 0.002 (227) – 118 ± 26 (227)	6.0 ± 0.2 (99)	0.09 ± 0.003 (583)	0.53 ± 0.008 (99)	0.049 ± 0.003 (583) – 124 ± 29 (583)
Novobiocin	1.71 ± 0.11 (86)	0.06 ± 0.003 (147)	0.11 ± 0.008 (86)	0.044 ± 0.004 (147) – 119 ± 35 (147)	5.5 ± 0.2 (153)	0.09 ± 0.003 (853)	0.52 ± 0.006 (153)	0.052 ± 0.002 (853) – 129 ± 27 (853)

Figure 7.24: Pre-rRNA cluster and RNAP cluster characteristics in pre-rRNA-RNAP two-color superresolution imaging experiments (fixed cell).

Pre-rRNA cluster colocalization values with RNAP clusters in fixed cell superresolution images under various conditions. Values shown are mean ± standard error. All colocalization values used a 50 nm distance threshold with (N) as the number of data points used in measurements. All simulations used the same N in the corresponding experimental data.

Condition	Fraction of rRNA clusters colocalizing to RNAP cluster	Fraction of rRNA clusters randomly colocalizing to RNAP clusters (simulation)	Fraction of RNAP clusters colocalizing to rRNA clusters	Fraction of RNAP clusters randomly colocalizing to rRNA clusters (simulation)
WT	0.69 ± 0.03 (720)	0.32 ± 0.03	0.83 ± 0.02 (404)	0.42 ± 0.02
Δ6rrn	0.68 ± 0.04 (586)	0.35 ± 0.03	0.76 ± 0.03 (247)	0.40 ± 0.03
Nalidixic acid	0.60 ± 0.05 (476)	0.39 ± 0.04	0.77 ± 0.04 (183)	0.46 ± 0.04
Novobiocin	0.52 ± 0.05 (378)	0.33 ± 0.04	0.79 ± 0.04 (148)	0.49 ± 0.05

Figure 7.25: Pre-rRNA cluster colocalization values with RNAP clusters in fixed cell superresolution images under various conditions.

7.11.2 Sample preparation and imaging conditions

A gel pad made with 3% low-melting-temperature agarose (SeaPlaque, Lonza) in the same growth media (or PBS for fixed cells) was prepared. Live cells were spun-down in a bench-top centrifuge at 8000 rpm for 2 min and resuspended in around 50 μl of fresh growth media (or PBS for fixed cells). An aliquot of 1 μl of the resuspension was then deposited to the agarose gel pad and cells immobilized between the gel pad and a coverslip for imaging as previously described [357, 358]. For fixed cell experiments, cells were fixed in 3.7% (v/v) paraformaldehyde (16% Paraformaldehyde, EM Grade, EMS) for 15 min at RT, washed with 1X PBS and imaged immediately. An Olympus IX-81 inverted microscope with a 100X oil objective (UPlanApo, N = 1.4x) was used, with 1.6x additional amplification. Images were captured with an Ixon DU-895 (Andor) EM-CCD with a 13 μm pixel size using MetaMorph (Molecular Devices). Illuminations (405 nm, 488 nm, 561 nm, 647 nm) were provided by solid-state lasers Coherent OBIS-405, Coherent OBIS-488, Coherent Sapphire-561, and Coherent OBIS-647 respectively. Fluorescence was split using a multi dichroic filter (ZT 405/488/561/647rpc, Chroma), and the far-red, red and green channels were further selected using HQ705/55, HQ600/50 and ET525/50 bandpass filters (Chroma). For two-color imaging, the simultaneous, multi-color acquisition was achieved using Optosplit II or Optosplit III (Cairn Research), colored channels were overlaid using calibration images from TetraSpeck beads (Life Technologies, T-7279) as previously described [356, 359], with around 10 nm registration error. Gold fiducial beads (50 nm, Microspheres-Nanospheres, Mahopac, NY) were used to correct for any sample

drift during imaging as previously described [139, 360]. All superresolution images were acquired with a 10 ms exposure time with 3000-9000 frames. Activation of fluorescent proteins was done simultaneously to fluorophore excitation, and activation laser was kept at a consistent power throughout the imaging session.

7.11.3 Superresolution imaging data analysis

Molecule localization and fitting of superresolution imaging data were done via thunderSTORM plugin (ImageJ, National Institutes of Health, Bethesda, MD) [204]. Subsequent analysis of localizations was performed using custom Matlab routines. See sections below for data analysis specifics.

7.11.4 Blinking correction

To correct for fluorophore blinking and its contribution to false clustering in superresolution images, we utilized a methodology we recently developed, Distance Distribution Correction (DDC) [313]. Briefly, DDC utilizes the finding that the pairwise distance distribution of localizations separated by a frame difference greater than the maximum lifetime of the fluorophore converges upon that of the "true localizations" (not due to the blinking of the emitters). DDC obtains a blinking corrected image by performing a phase search, eliminating localizations of high blinking probabilities, so that the pairwise distance distribution at all frames is consistent with the "true pairwise distance distribution." We verified that this methodology was significantly more

accurate compared to the commonly used thresholding methodology using a variety of simulations and diverse clustering structures, providing the most rigorously scrutinized representation for the locations of the underlying molecules to date.

7.11.5 Cluster identification

To determine a cluster across the different experimental conditions and different molecular species, we normalized the number of localizations by cell volume so that each cell had the same concentration of localizations. The concentration normalization eliminated the effects of cell size and the noise in detection efficiency from being the dominating factors in the characteristics of the clusters. To do this, we first calculated the volume by outlining the shape of each cell using the outermost localizations to determine an area; this area was then projecting to a 3D volume. We only used cells that had enough localizations (at least 700 localizations) to reach the desired concentration threshold for each species.

To obtain the properties of individual clusters, for each species in various conditions, we first eliminated localizations in low-density areas. By calculating the average distance to the closest ten localizations surrounding each localization, we determined whether each localization was in a high-density region if the average distance was greater than a specified value. This calculation was only valid since each cell had the same concentration of molecules, which allowed us to use one defined threshold for each species.

Including only localizations within the high-density regions, we applied a

tree-cluster algorithm in MATLAB. Specifically, we utilized the 'single' method using the linkage function, which provided us with a tree of hierarchical clusters for the data. We then used the cluster function with a cutoff of 100 nm and the distance criterion. The analysis linked all localizations together as one cluster if they are within 100 nm of each other. As a final step, we only counted clusters that possessed more than a certain percentage of the total localizations. All of the analysis code are available via GitHub [361].

7.11.6 Random distribution simulation

To determine whether the clustering of a species was significant, a random distribution of localizations was created, analyzed and compared for each species and condition. To simulate the random distribution of localizations, we first determined the volume of each cell for each condition (as discussed in the cluster determination section). We then randomly placed localizations within this 3D volume according to a uniform distribution; the number of localizations used closely matched to the experimentally collected molecule number for each cell. We then adjusted the concentration of molecules to match the desired concentration used in the cluster determination section and applied the same clustering analyses.

7.11.7 Colocalization

We calculated the colocalization value from one species' cluster to the other species' clusters as the following. For a cluster of species A (for example A1), and the clusters of species B (B1 to Bn) in the same cell, we calculated

the pairwise distances between the localizations in (A1) to the localizations in any B (B1 to Bn) and recorded the shortest distance (d1). We repeated this calculation for all the other clusters of species A (A1 to An) in the same cell. Therefore, each cluster of A (A1 to An) in the cell was associated with a distance (d1, d2 to dj). Next, we repeated the same calculation for all clusters of species A in all cells and obtained a data structure in which a cluster i of species A in cell m A_i^m has a distance d_i^m. We then selected a threshold distance and counted the number (n) of clusters of species A that had at least one distance shorter or equal to the threshold distance. Note that we only performed this calculation if both species had clusters within the same cell. The colocalization value of clusters of species A to clusters of species B was calculated by dividing this number n by the total number (N) of clusters of species A and plotted as a cumulative curve at different threshold distances. As such, a colocalization fraction of 0.8 of RNAP clusters to pre-rRNA clusters at 50 nm means that at 50 nm, 80% of RNAP clusters had at least one pre-rRNA cluster within a distance of 50 nm. Beside the cumulative curves, we also reported colocalization values at a set distance threshold for all experiments conducted in this work (Methods, Fig. 7.25) for ease of comparison between different conditions. Note that the colocalization value from species A to species B is different from the reverse direction (from species B to species A) and we reported both in Methods, Fig. 7.25.

7.11.8 Accounting for experimental cluster detection efficiency

In calculating the colocalization value between two species' clusters, it is important to consider the detection efficiency of each species' clusters. Assuming that the detection efficiency for species B is p ($p < 1$), and that the true colocalization value from species A to species B is q, the measured colocalization value c from species A to B will then be modified by the detection efficiency p as c=p*q. Therefore, the true colocalization value q should be calculated as q=c/p.

To determine the detection efficiency of pre-rRNA clusters for the rich medium growth condition, we used two L1 probes with the same sequence but different dye labels (Alexa Fluor 488 and 647, respectively) to hybridize with pre-rRNAs in the same cells. Because the probe sequences were the same, pre-rRNA clusters identified by the two colors should be identical and colocalize with each other 100%. Therefore, the lower than 100% colocalization value we detected from one color to the other, likely resulting from dye properties and cluster thresholding, allowed us to calculate the detection efficiency of the L1 probe. As shown in Methods, Fig. 7.9A, we observed nearly identical cumulative curves (After blinking correction) of the colocalization value from L1-Alexa Fluor 488 to L1-Alexa Fluor 647 and vice versa. At 50 nm, the detection efficiency of both probes was approximately 80%. To determine the detection efficiency of RNAP clusters, we used a computational approach (Methods, Fig. 7.10) due to the inability to obtain a fully functional RNAP-Dronpa-PAmCherry tandem dimer fusion on the chromosome. In the compu-

tational approach, we randomly split into two channels RNAP localizations of cells that had at least twice the predefined concentration of localizations, so that there were two sets of localizations in a cell with the desired concentration, mimicking cluster detection using two different colors. We then performed the cluster analysis on each set of localizations and calculated the colocalization value between the clusters identified in the two sets at different threshold distances (Methods, Fig. 7.10). We further verified this computational approach using the experimentally measured colocalization curves of the L1 probes of two different dyes and obtained the same result (Methods, Fig. 7.9A).

The colocalization cumulative curves between the two sets of clusters provided us with the best possible colocalization at each distance given our detection efficiency. In all colocalization curves reported in this work except for Methods, Fig. 7.9A and 7.10, we adjusted the colocalization values by dividing the measured colocalization value by the measured detection efficiency value at the same distance.

7.11.9 smFISH - L1 probe labeling of pre-rRNA

We performed smFISH using a previously published protocol [357, 200]. Briefly, cells were grown in EZRDM glucose as previously described; 5 ml of mid-log phase cells were fixed with 3.7% (v/v) paraformaldehyde (16% Paraformaldehyde, EM Grade, EMS), placed for 30 min on ice. Next, cells were harvested via centrifugation, and subsequently washed two times in 1X PBS. Cells were then permeabilized by resuspending in a mixture of 300 μl of

H2O and 700 μl of 100% ethanol and incubating with rotation at RT for 30 min. Cells were stored at 4C until next day. Wash buffer was freshly prepared with 40% formamide and 2x SSC and put on ice. Cells were spun-down in a bench-top centrifuge at 10000 rpm for 3 min and the cell pellet was resuspended in 1 ml of wash buffer. The sample was placed on a nutator to mix for 5 min at RT. Hybridization solution was prepared with 40% formamide and 2x SSC, subsequently, dye-labeled oligo probes were added to hybridization solution to a final concentration of 1 μM. Cells were spun-down again and 50 μl of hybridization solution with probe was added to the pellet. The hybridization sample was mixed well and placed overnight in a 30°C incubator. Next day, 10 μl of hybridization sample was washed with 200 μl of fresh wash buffer and incubated at 30°C for 30 min, this was repeated one more time. The washed sample was imaged immediately: without STORM imaging buffer for ensemble fluorescence, with STORM buffer to induce dye blinking for superresolution imaging. glucose oxidase + thiol STORM buffer was used to image samples with only dye labeling (50 mM Tris (pH 8.0), 10 mM NaCl, 0.5 mg ml^{-1} glucose oxidase (Sigma-Aldrich), 40 μg ml^{-1} catalase (Roche), 10% (w/v) glucose and 10 mM MEA (Fluka)) [201]. Thiol only STORM buffer (10 mM MEA, 50 mM Tris (pH 8.0), 10 mM NaCl) was used to image samples with both endogenously expressed fluorescent proteins and dye labeling. This was to preserve the fluorescent signal from fluorescent proteins, since the presence of glucose oxidase in the STORM buffer tended to quench the fluorescent protein signal.

Pre-rRNA transcripts were detected with a single probe L1, conjugated at

the 5' with either Alexa Fluor 488 (NHS ester) or Alexa Fluor 647 (NHS ester) (IDT) (Methods, Fig. 7.6A) [195]. Upon receiving the commercial oligos, a working stock (50 μM) was made and aliquoted for storage at -20C. Image processing of smFISH ensemble fluoresence images:

Ensemble intensity measurements were performed using ImageJ (National Institutes of Health, Bethesda, MD). Ensemble fluorescence images with focus plane at mid-cell were segmented manually using their corresponding bright-field images. Each cell's total fluorescent intensity was calculated as: (area of the segmented cell x (mean intensity inside cell - mean intensity of background region)). Around 50-100 cells were used to represent the total cellular fluoresence intensities for a single experimental condition. See corresponding figure captions for the exact number of cells used in the calculations.

7.11.10 DNA staining in fixed cells using Hoechst dye (33342)

Hoechst dye (bisBenzimide H33342 trihydrochloride) was used to stain chromosomal DNA in *E. coli* cells. Stained cells were subsequently visualized via 3D SIM on a GE OMX SR SIM scope, with a 60x objective. Briefly, cells were grown to mid-log phase (O.D.600 = 0.3-0.4) in EZRDM at RT. For all conditions except for Δ6rrn, the strain RpoC-PAmCherry was used for DNA staining and considered as the wt strain. Hoechst dye (0.5 μl of 10 mg ml^{-1} stock) was added during the last 10 min of cell growth. After 10 min of incubation with Hoechst dye, 1 ml of the liquid culture was immediately harvested and spun-down in a bench-top centrifuge at 8000 rpm for 2 min and the cell pellet was washed with 1 ml of 1x PBS. For fixation, the cell pellet was

resuspended in 1 ml of 3% paraformaldehyde in 1x PBS, placed on a nutator and fixed at RT for 15 min. After fixation, cells were subsequently spun-down at 8000 rpm for 2 min and the cell pellet was washed with 1 ml of 1x PBS. About 35 μl of 1x PBS was used to resuspend the cell pellet as a final step before mounting. An equal volume of fixed cells in PBS and anti-fading buffer (60% glycerol, 20% NPG (n-propyl gallate, 1x PBS) was combined to a total of 50 μl and used for mounting between Poly-L-Lysine treated coverslip and cover glass. Excess liquid was siphoned away with a Kimwipe tissue, and the coverslip was sealed on the glass slide using clear nail polish. Imaging was performed 30-60 min after sample preparation.

3D SIM Imaging conditions were as follows: 5% laser power for 405 nm laser excitation, 30 ms exposure time, with standard 125-nm interval Z-sections, and a pixel size of 40 nm. Images were collected using the standard SRx software and reconstructed using standard SIM reconstruction parameters.

For a more quantitative comparison between different experimental conditions, we calculated the nucleoid territory occupancy in cells. Briefly, we used intensity thresholding to isolate both the cell area voxels (a lower intensity threshold), and the DNA area voxels (a higher intensity threshold) and used (DNA area/cell area) to calculated the percentage of total cell area that was occupied by DNA, a representative 15 cells were used for this calculation for each experimental condition. Additionally, we constructed 2D histograms of the DNA signal intensities for each condition. The DNA intensity from the projected Z-stack of the eight 125-nm Z slices for each cell were combined for each condition, 15 cells were used for each experimental condition, the cells

were rotated, and the long axis was normalized to each cell's cell length; the 2D histograms were represented in a standard 1 μm x 3 μm cell.

7.11.11 Co-immunoprecipitation and western blot

Co-immunoprecipitation was performed using Protein G Sepharose beads (Abcam, ab193259) following the manufacture-provided protocol. Briefly, 2 L of cultured *E. coli* cells grown to O.D=0.3-0.5 in LB were spun down and the pellet was resuspended in 50 ml of 2x lysis buffer (20 mM Tris-HCL pH8, 150 mM KCl, 1 mM DTT, 100 mM PMSF, PIC cocktail, 10 mg/ml lysozyme, and 10 units of DNase I, NEB M0303S). This mixture was frozen at -80C for at least 1 hr to overnight. After 30 min thawed at RT, the lysis solution was vortexed with short rests on ice to help facilitate cell lysis. This lysis solution was then spun down at 15,000 rpm for 30 min at 4C to pellet cells. The lysis solution supernatant was decanted into a new tube, and saved on ice until later steps; the cell pellet was discarded. In the meanwhile, for each sample, 100 μl of fresh bead slurry was prepared by washing in 500 μl of dilution buffer (10 mM Tris-HCl, 150 mM NaCl, 0.5 mM EDTA) three times and the final bead slurry volume was kept at 100 μl. Then to this 100 μl of washed bead slurry, 1 μl of the cell lysate (from the 50 ml total cell lysate) in 499 μl of dilution buffer along with 5 μl of anti-RpoB capture antibody (Mouse IgG1 BioLegend Cat. 663903) was added, and this mixture (approximately 600 μl) was incubated and nutated for 2 hrs at 4C. The mixture was then spun down and 500 μl beads flow-through was removed and stored at 4C to be run on SDS-PAGE together with bead elute as described below. The bead slurry left (approximately 100 μl) was washed once in 500 μl of dilution buffer, pelleted, and then incubated

with 100 μl of 2x SDS buffer at 4C overnight. The next day, the beads in the 2X SDS buffer was boiled for 5 min and a fraction (5 or 10 μl) of the total eluted supernatant (100 μl) was loaded to a 4-15% SDS-PAGE gel to run at 180V for 45 min together with the same volume from the 500 μl beads flow-through. The SDS-PAGE gel was visualized via western blot using antibodies against mCherry (Rabbit, ab167453 which also targets PAmCherry, and goat anti-rabbit HRP was used along with the Clarity Western ECL Substrate (BioRad, #1705061). The blot was imaged using x-ray film.

We measured the band intensity in the western blot using image J and compensated for the difference in dilution between the beads flow-through vs. the eluate, which was 500 μl : 100 μl, or 5 : 1. We multiplied the band intensity of the beads flow-through lane by 5, and determined the percentage of incorporation of the RpoC-PAmCherry subunit into the core/holoenzyme. We performed the same experiment twice and obtained the percentage at 85.9% and 90.7% with the average at 88.3%.

7.11.12 Cell length determination

Cells were grown as previously described and mounted onto agarose pads. Brightfield images of cells were taken. Individual cells were cropped and their lengths were manually measured between the two cell poles (see Methods, Fig. 7.16). Individual cells from different samples were blinded to prevent bias from the manual measurement.

Bibliography

[1] Rayan, G., J.-E. Guet, N. Taulier, F. Pincet, and W. Urbach, 2010. Recent Applications of Fluorescence Recovery after Photobleaching (FRAP) to Membrane Bio-Macromolecules. Sensors 10:5927–5948.

[2] Elowitz, M. B., M. G. Surette, P. E. Wolf, J. B. Stock, and S. Leibler, 1999. Protein mobility in the cytoplasm of Escherichia coli. Journal of bacteriology 181:197–203.

[3] Elson, E. L., 2011. Fluorescence Correlation Spectroscopy: Past, Present, Future. Biophysj 101:2855–2870.

[4] Michalet, X. and A. J. Berglund, 2012. Optimal diffusion coefficient estimation in single-particle tracking. Physical review. E, Statistical, nonlinear, and soft matter physics 85:061916.

[5] Betzig, E., G. H. Patterson, R. Sougrat, O. W. Lindwasser, S. Olenych, J. S. Bonifacino, M. W. Davidson, J. Lippincott-Schwartz, and H. F. Hess, 2006. Imaging intracellular fluorescent proteins at nanometer resolution. Science 313:1642–1645.

[6] Rust, M. J., M. Bates, and X. Zhuang, 2006. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). Nature Methods 3:793–796.

[7] Hess, S. T., T. P. K. Girirajan, and M. D. Mason, 2006. Ultra-high resolution imaging by fluorescence photoactivation localization microscopy. Biophysj 91:4258–4272.

[8] Das, R., C. W. Cairo, and D. Coombs, 2009. A hidden Markov model for single particle tracks quantifies dynamic interactions between LFA-1 and the actin cytoskeleton. PLoS Computational Biology 5:e1000556.

[9] Bohrer, C. H., K. Bettridge, and J. Xiao, 2017. Reduction of Confinement Error in Single-Molecule Tracking in Live Bacterial Cells Using SPICER. Biophysical journal 112:568–574.

[10] Persson, F., M. Lindén, C. Unoson, and J. Elf, 2013. Extracting intracellular diffusive states and transition rates from single-molecule tracking data. Nature Methods 10:265–269.

[11] Slator, P. J. and N. J. Burroughs, 2018. A Hidden Markov Model for Detecting Confinement in Single-Particle Tracking Trajectories. Biophysical journal 115:1741–1754.

[12] Matsuda, Y., I. Hanasaki, R. Iwao, H. Yamaguchi, and T. Niimi, 2018. Estimation of diffusive states from single-particle trajectory in heterogeneous medium using machine-learning methods. Physical chemistry chemical physics : PCCP 20:24099–24108.

[13] Weber, S. C., M. A. Thompson, W. E. Moerner, A. J. Spakowitz, and J. A. Theriot, 2012. Analytical tools to distinguish the effects of localization error, confinement, and medium elasticity on the velocity autocorrelation function. Biophysical journal 102:2443–2450.

[14] Weber, S. C., A. J. Spakowitz, and J. A. Theriot, 2010. Bacterial chromosomal loci move subdiffusively through a viscoelastic cytoplasm. Physical Review Letters 104:238102.

[15] He, Y., S. Burov, R. Metzler, and E. Barkai, 2008. Random time-scale invariant diffusion and transport coefficients. Physical Review Letters 101:058101.

[16] Condamin, S., V. Tejedor, R. Voituriez, O. Bénichou, and J. Klafter, 2008. Probing microscopic origins of confined subdiffusion by first-passage observables. Proceedings of the National Academy of Sciences of the United States of America 105:5675–5680.

[17] Thapa, S., M. A. Lomholt, J. Krog, A. G. Cherstvy, and R. Metzler, 2018. Bayesian analysis of single-particle tracking data using the nested-sampling algorithm: maximum-likelihood model selection applied to stochastic-diffusivity data. Physical chemistry chemical physics : PCCP 20:29018–29037.

[18] Martin, D. S., M. B. Forstner, and J. A. Käs, 2002. Apparent Subdiffusion Inherent to Single Particle Tracking. Biophysical journal 83:2109–2117.

[19] Lee, S.-H., J. Y. Shin, A. Lee, and C. Bustamante, 2012. Counting single photoactivatable fluorescent molecules by photoactivated localization microscopy (PALM). Proceedings of the National Academy of Sciences of the United States of America 109:17436–17441.

[20] Edwin K L Yeow, Sergey M Melnikov, Toby D M Bell, F. C. D. Schryver, , and J. Hofkens, 2006. Characterizing the Fluorescence Intermittency and Photobleaching Kinetics of Dye Molecules Immobilized on a Glass Surface. The Journal of Physical Chemistry A 110:1726–1734.

[21] Deng, W. and E. Barkai, 2009. Ergodic properties of fractional Brownian-Langevin motion. Physical Review E 79:011112.

[22] Parry, B. R., I. V. Surovtsev, M. T. Cabeen, C. S. O'Hern, E. R. Dufresne, and C. Jacobs-Wagner, 2014. The Bacterial Cytoplasm Has Glass-like Properties and Is Fluidized by Metabolic Activity. Cell 156:183–194.

[23] Lampo, T. J., S. Stylianidou, M. P. Backlund, P. A. Wiggins, and A. J. Spakowitz, 2017. Cytoplasmic RNA-Protein Particles Exhibit Non-Gaussian Subdiffusive Behavior. Biophysj 112:532–542.

[24] Lukinavičius, G., K. Umezawa, N. Olivier, A. Honigmann, G. Yang, T. Plass, V. Mueller, L. Reymond, I. R. Corrêa Jr, Z.-G. Luo, C. Schultz, E. A. Lemke, P. Heppenstall, C. Eggeling, S. Manley, and K. Johnsson, 2013. A near-infrared fluorophore for live-cell super-resolution microscopy of cellular proteins. Nature Chemistry 5:132–139.

[25] Los, G. V., L. P. Encell, M. G. McDougall, D. D. Hartzell, N. Karassina, C. Zimprich, M. G. Wood, R. Learish, R. F. Ohana, M. Urh, D. Simpson, J. Mendez, K. Zimmerman, P. Otto, G. Vidugiris, J. Zhu, A. Darzins, D. H. Klaubert, R. F. Bulleit, and K. V. Wood, 2008. HaloTag: A Novel Protein Labeling Technology for Cell Imaging and Protein Analysis. ACS chemical biology 3:373–382.

[26] Cole, N. B., 2013. Site-Specific Protein Labeling with SNAP-Tags. Current Protocols in Protein Science 73:30.1.1–30.1.16.

[27] Xiao, J., 2009. Single-Molecule Imaging in Live Cells. In Handbook of Single-Molecule Biophysics, Springer, New York, NY, New York, NY, 43–93.

[28] Beilharz, K., R. van Raaphorst, M. Kjos, J.-W. Veening, and M. J. Pettinari, 2015. Red Fluorescent Proteins for Gene Expression and Protein Localization Studies in Streptococcus pneumoniae and Efficient Transformation with DNA Assembled via the Gibson Assembly Method. Applied and Environmental Microbiology 81:7244–7252.

[29] Zhang, G., V. Gurtu, and S. R. Kain, 1996. An Enhanced Green Fluorescent Protein Allows Sensitive Detection of Gene Transfer in Mammalian Cells. Biochemical and biophysical research communications 227:707–711.

[30] Ormö, M., A. B. Cubitt, K. Kallio, L. A. Gross, R. Y. Tsien, and S. J. Remington, 1996. Crystal Structure of the Aequorea victoria Green Fluorescent Protein. Science 273:1392–1395.

[31] Shaner, N. C., R. E. Campbell, P. A. Steinbach, B. N. G. Giepmans, A. E. Palmer, and R. Y. Tsien, 2004. Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein. Nature biotechnology 22:1567–1572.

[32] Shaner, N. C., G. G. Lambert, A. Chammas, Y. Ni, P. J. Cranfill, M. A. Baird, B. R. Sell, J. R. Allen, R. N. Day, M. Israelsson, M. W. Davidson, and J. Wang, 2013. A bright monomeric green fluorescent protein derived from Branchiostoma lanceolatum. Nature Methods 10:407–409.

[33] Zhang, M., H. Chang, Y. Zhang, J. Yu, L. Wu, W. Ji, J. Chen, B. Liu, J. Lu, Y. Liu, J. Zhang, P. Xu, and T. Xu, 2012. Rational design of true monomeric and bright photoactivatable fluorescent proteins. Nature Methods 9:727–729.

[34] Subach, F. V., G. H. Patterson, S. Manley, J. M. Gillette, J. Lippincott-Schwartz, and V. V. Verkhusha, 2009. Photoactivatable mCherry for high-resolution two-color fluorescence microscopy. Nature Methods 6:153–159.

[35] Gebhardt, J. C. M., D. M. Suter, R. Roy, Z. W. Zhao, A. R. Chapman, S. Basu, T. Maniatis, and X. S. Xie, 2013. Single-molecule imaging of transcription factor binding to DNA in live mammalian cells. Nature Methods 10:421–426.

[36] Stepanenko, O. V., O. V. Stepanenko, D. M. Shcherbakova, I. M. Kuznetsova, K. K. Turoverov, and V. V. Verkhusha, 2011. Modern fluorescent proteins: from chromophore formation to novel intracellular applications. BioTechniques 51:313–4– 316– 318 passim.

[37] Ando, R., H. Hama, M. Yamamoto-Hino, H. Mizuno, and A. Miyawaki, 2002. An optical marker based on the UV-induced green-to-red photoconversion of a fluorescent protein. Proceedings of the National Academy of Sciences 99:12651–12656.

[38] Gurskaya, N. G., V. V. Verkhusha, A. S. Shcheglov, D. B. Staroverov, T. V. Chepurnykh, A. F. Fradkov, S. Lukyanov, and K. A. Lukyanov, 2006. Engineering of a monomeric green-to-red photoactivatable fluorescent protein induced by blue light. Nature biotechnology 24:461–465.

[39] Grimm, J. B., B. P. English, H. Choi, A. K. Muthusamy, B. P. Mehl, P. Dong, T. A. Brown, J. Lippincott-Schwartz, Z. Liu, T. Lionnet, and L. D. Lavis, 2016. Bright photoactivatable fluorophores for single-molecule imaging. Nature Methods 13:985–988.

[40] Stracy, M., C. Lesterlin, F. Garza de Leon, S. Uphoff, P. Zawadzki, and A. N. Kapanidis, 2015. Live-cell superresolution microscopy reveals the organization of RNA polymerase in the bacterial nucleoid. Proceedings of the National Academy of Sciences of the United States of America 112:E4390–E4399.

[41] Lampo, T. J., S. Stylianidou, M. P. Backlund, P. A. Wiggins, and A. J. Spakowitz, 2017. Cytoplasmic RNA-Protein Particles Exhibit Non-Gaussian Subdiffusive Behavior. Biophysj 112:532–542.

[42] Vrljic, M., S. Y. Nishimura, and W. E. Moerner, 2007. Single-molecule tracking. Methods in molecular biology (Clifton, N.J.) 398:193–219.

[43] Fu, G., J. N. Bandaria, A. V. Le Gall, X. Fan, A. Yildiz, T. Mignot, D. R. Zusman, and B. Nan, 2018. MotAB-like machinery drives the movement of MreB filaments during bacterial gliding motility. Proceedings of the National Academy of Sciences of the United States of America 115:2484–2489.

[44] Yang, X., Z. Lyu, A. Miguel, R. McQuillen, K. C. Huang, and J. Xiao, 2017. GTPase activity-coupled treadmilling of the bacterial tubulin FtsZ organizes septal cell wall synthesis. Science 355:744–747.

[45] Perez, A. J., Y. Cesbron, S. L. Shaw, J. Bazan Villicana, H.-C. T. Tsui, M. J. Boersma, Z. A. Ye, Y. Tovpeko, C. Dekker, S. Holden, and M. E. Winkler, 2019. Movement dynamics of divisome proteins and PBP2x:FtsW in cells of Streptococcus pneumoniae. Proceedings of the National Academy of Sciences of the United States of America 116:3211–3220.

[46] Bisson-Filho, A. W., Y.-P. Hsu, G. R. Squyres, E. Kuru, F. Wu, C. Jukes, Y. Sun, C. Dekker, S. Holden, M. S. VanNieuwenhze, Y. V. Brun, and E. C. Garner, 2017. Treadmilling by FtsZ filaments drives peptidoglycan synthesis and bacterial cell division. Science 355:739–743.

[47] Ringgaard, S., J. van Zon, M. Howard, and K. Gerdes, 2009. Movement and equipositioning of plasmids by ParA filament disassembly. Proceedings of the National Academy of Sciences of the United States of America 106:19369–19374.

[48] Hu, L., A. G. Vecchiarelli, K. Mizuuchi, K. C. Neuman, and J. Liu, 2015. Directed and persistent movement arises from mechanochemistry of the ParA/ParB system. Proceedings of the National Academy of Sciences of the United States of America 112:E7055–64.

[49] Kim, S. Y., Z. Gitai, A. Kinkhabwala, L. Shapiro, and W. E. Moerner, 2006. Single molecules of the bacterial actin MreB undergo directed treadmilling motion in Caulobacter crescentus. Proceedings of the National Academy of Sciences 103:10929–10934.

[50] Kusumi, A., Y. Sako, and M. Yamamoto, 1993. Confined lateral diffusion of membrane receptors as studied by single particle tracking (nanovid microscopy). Effects of calcium-induced differentiation in cultured epithelial cells. Biophysical journal 65:2021–2040.

[51] Bakshi, S., R. M. Dalrymple, W. Li, H. Choi, and J. C. Weisshaar, 2013. Partitioning of RNA Polymerase Activity in Live Escherichia coli from Analysis of Single-Molecule Diffusive Trajectories. Biophysical journal 105:2676–2686.

[52] Havlin, S., D. B.-A. A. i. Physics, and 1987, 1987. Diffusion in disordered media. Taylor & Francis 36:695–798.

[53] Hunter, G. L. and E. R. Weeks, 2012. The physics of the colloidal glass transition. Reports on progress in physics. Physical Society (Great Britain) 75:066501.

[54] Weeks, E. R. and D. A. Weitz, 2002. Properties of cage rearrangements observed near the colloidal glass transition. Physical Review Letters 89:095704.

[55] Cipelletti, L. and L. Ramos, 2005. Slow dynamics in glassy soft matter. Journal of Physics: Condensed Matter 17:R253–R285.

[56] Balakrishnan, V., 1985. Anomalous diffusion in one dimension. Physica A: Statistical Mechanics and its Applications 132:569–580.

[57] Weber, S. C., J. A. Theriot, and A. J. Spakowitz, 2010. Subdiffusive motion of a polymer composed of subdiffusive monomers. Physical Review E 82:011913.

[58] Weiss, M., 2013. Single-particle tracking data reveal anticorrelated fractional Brownian motion in crowded fluids. Physical Review E 88:010101.

[59] Lutz, E., 2001. Fractional Langevin equation. Physical review. E, Statistical, nonlinear, and soft matter physics 64:051106.

[60] Weber, S. C., A. J. Spakowitz, and J. A. Theriot, 2012. Nonthermal ATP-dependent fluctuations contribute to the in vivo motion of chromosomal loci. Proceedings of the National Academy of Sciences of the United States of America 109:7338–7343.

[61] Pan, W., L. Filobelo, N. D. Q. Pham, O. Galkin, V. V. Uzunova, and P. G. Vekilov, 2009. Viscoelasticity in Homogeneous Protein Solutions. Physical Review Letters 102:108–4.

[62] Nagle, J. F., 1992. Long tail kinetics in biophysics? Biophysj 63:366–370.

[63] Cayley, S., B. A. Lewis, H. J. Guttman, and M. T. Record Jr, 1991. Characterization of the cytoplasm of Escherichia coli K-12 as a function of external osmolarity: Implications for protein-DNA interactions in vivo. Journal of molecular biology 222:281–300.

[64] Winick, M., 1968. Changes in Nucleic Acid and Protein Content of the Human Brain During Growth. Pediatric Research 2:352–355.

[65] Mika, J. T. and B. Poolman, 2011. Macromolecule diffusion and confinement in prokaryotic cells. Current opinion in biotechnology 22:117–126.

[66] Terry, B. R., E. K. Matthews, and J. Haseloff, 1995. Molecular characterisation of recombinant green fluorescent protein by fluorescence correlation microscopy. Biochemical and biophysical research communications 217:21–27.

[67] Swaminathan, R., C. P. Hoang, and A. S. Verkman, 1997. Photobleaching recovery and anisotropy decay of green fluorescent protein GFP-S65T in solution and cells: cytoplasmic viscosity probed by green fluorescent protein translational and rotational diffusion. Biophysj 72:1900–1907.

[68] Einstein, A., 2007. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen :1–12.

[69] Kumar, M., M. S. Mommer, and V. Sourjik, 2010. Mobility of cytoplasmic, membrane, and DNA-binding proteins in Escherichia coli. Biophysical journal 98:552–559.

[70] Golding, I. and E. C. Cox, 2006. Physical Nature of Bacterial Cytoplasm. Physical Review Letters 96:098102.

[71] Konopka, M. C., I. A. Shkel, S. Cayley, M. T. Record, and J. C. Weisshaar, 2006. Crowding and confinement effects on protein diffusion in vivo. Journal of bacteriology 188:6115–6123.

[72] van den Bogaart, G., N. Hermans, V. Krasnikov, and B. Poolman, 2007. Protein mobility and diffusive barriers in Escherichia coli: consequences of osmotic stress. Molecular microbiology 64:858–871.

[73] Winther, T., L. Xu, K. Berg-Soslashrensen, S. Brown, and L. B. Oddershede, 2009. Effect of Energy Metabolism on Protein Motility in the Bacterial Outer Membrane. Biophysj 97:1305–1312.

[74] Sadoon, A. A. and Y. Wang, 2018. Anomalous, non-Gaussian, viscoelastic, and age-dependent dynamics of histonelike nucleoid-structuring proteins in live Escherichia coli :1–8.

[75] Schavemaker, P. E., W. M. Śmigiel, and B. Poolman, 2017. Ribosome surface properties may impose limits on the nature of the cytoplasmic proteome. eLife 6.

[76] Drlica, K. and J. Rouviere-Yaniv, 1987. Histonelike proteins of bacteria. Microbiological Reviews 51:301–319.

[77] Stracy, M., A. J. Wollman, E. Kaja, J. Gapinski, J.-E. Lee, V. A. Leek, S. J. McKie, L. A. Mitchenall, A. Maxwell, D. J. Sherratt, M. C. Leake, and P. Zawadzki, 2018. Single-molecule imaging of DNA gyrase activity in living Escherichia coli. Nucleic acids research 6:11055–11.

[78] Fudenberg, G., M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur, and L. A. Mirny, 2016. Formation of Chromosomal Domains by Loop Extrusion. Cell reports 15:2038–2049.

[79] Dorman, C. J. and M. J. Dorman, 2016. DNA supercoiling is a fundamental regulatory principle in the control of bacterial gene expression. Biophysical Reviews 8:89–100.

[80] Postow, L., C. D. Hardy, J. Arsuaga, and N. R. Cozzarelli, 2004. Topological domain structure of the Escherichia coli chromosome. Genes & Development 18:1766–1779.

[81] Bohrer, C. H. and E. Roberts, 2016. A biophysical model of supercoiling dependent transcription predicts a structural aspect to gene regulation. BMC biophysics 9:1.

[82] Chong, S., C. Chen, H. Ge, and X. S. Xie, 2014. Mechanism of Transcriptional Bursting in Bacteria. Cell 158:314–326.

[83] Weng, X., C. H. Bohrer, K. Bettridge, A. C. Lagda, C. Cagliero, D. J. Jin, and J. Xiao, 2018. RNA polymerase organizes into distinct spatial clusters independent of ribosomal RNA transcription in E. coli. bioRxiv :320481.

[84] Kapanidis, A. N., S. Uphoff, and M. Stracy, 2018. Understanding Protein Mobility in Bacteria by Tracking Single Molecules. Journal of molecular biology 430:4443–4455.

[85] Nielsen, H. J., Y. Li, B. Youngren, F. G. Hansen, and S. Austin, 2006. Progressive segregation of the Escherichia coli chromosome. Molecular microbiology 61:383–393.

[86] von Hippel, P. H. and O. G. Berg, 1989. Facilitated target location in biological systems. Journal of Biological Chemistry 264:675–678.

[87] Hobot, J. A., W. Villiger, J. Escaig, M. Maeder, A. Ryter, and E. Kellenberger, 1985. Shape and fine structure of nucleoids observed on sections of ultrarapidly frozen and cryosubstituted bacteria. Journal of bacteriology 162:960–971.

[88] Wang, W., G.-W. Li, C. Chen, X. S. Xie, and X. Zhuang, 2011. Chromosome organization by a nucleoid-associated protein in live bacteria. Science 333:1445–1449.

[89] Bakshi, S., A. Siryaporn, M. Goulian, and J. C. Weisshaar, 2012. Superresolution imaging of ribosomes and RNA polymerase in live Escherichia coli cells. Molecular microbiology 85:21–38.

[90] Stracy, M., S. Uphoff, F. Garza de Leon, and A. N. Kapanidis, 2014. In vivo single-molecule imaging of bacterial DNA replication, transcription, and repair. FEBS Letters 588:3585–3594.

[91] Ruiz, N., D. Kahne, and T. J. Silhavy, 2006. Advances in understanding bacterial outer-membrane biogenesis. Nature reviews. Microbiology 4:57–66.

[92] Lessen, H. J., P. J. Fleming, K. G. Fleming, and A. J. Sodt, 2018. Building Blocks of the Outer Membrane: Calculating a General Elastic Energy Model for β-Barrel Membrane Proteins. Journal of Chemical Theory and Computation 14:4487–4497.

[93] Cho, S.-H., J. Szewczyk, C. Pesavento, M. Zietek, M. Banzhaf, P. Roszczenko, A. Asmar, G. Laloux, A.-K. Hov, P. Leverrier, C. Van der Henst, D. Vertommen, A. Typas, and J.-F. Collet, 2014. Detecting Envelope Stress by Monitoring β-Barrel Assembly. Cell 159:1652–1664.

[94] Goemans, C., K. Denoncin, and J.-F. Collet, 2014. Folding mechanisms of periplasmic proteins. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research 1843:1517–1528.

[95] Denoncin, K., D. Vertommen, I. S. Arts, C. V. Goemans, S. Rahuel-Clermont, J. Messens, and J.-F. Collet, 2014. A new role for Escherichia coli DsbC protein in protection against oxidative stress. Journal of Biological Chemistry 289:12356–12364.

[96] Mas, G., J. Thoma, and S. Hiller, 2019. The Periplasmic Chaperones Skp and SurA. In Bacterial Cell Walls and Membranes, Springer, Cham, Cham, 169–186.

[97] Grote, G. R. M. K., J. M. Risse, and K. Friehs, 2018. Secretion of recombinant proteins from E. coli. Engineering in Life Sciences 18:532–550.

[98] Foley, M., J. M. Brass, J. Birmingham, W. R. Cook, P. B. Garland, C. F. Higgins, and L. I. Rothfield, 1989. Compartmentalization of the periplasm at cell division sites in Escherichia coli as shown by fluorescence photobleaching experiments. Molecular microbiology 3:1329–1336.

[99] Mullineaux, C. W., A. Nenninger, N. Ray, and C. Robinson, 2006. Diffusion of green fluorescent protein in three cell environments in Escherichia coli. Journal of bacteriology 188:3442–3448.

[100] Sochacki, K. A., I. A. Shkel, M. T. Record, and J. C. Weisshaar, 2011. Protein Diffusion in the Periplasm of E. coli under Osmotic Stress. Biophysj 100:22–31.

[101] Rudner, D. Z. and R. Losick, 2010. Protein subcellular localization in bacteria. Cold Spring Harbor perspectives in biology 2:a000307–a000307.

[102] Lopez, D. and G. Koch, 2017. Exploring functional membrane microdomains in bacteria: an overview. Current opinion in microbiology 36:76–84.

[103] Dempwolff, F., F. K. Schmidt, A. B. Hervás, A. Stroh, T. C. Rösch, C. N. Riese, S. Dersch, T. Heimerl, D. Lucena, N. Hülsbusch, C. A. O. Stuermer, N. Takeshita, R. Fischer, B. Eckhardt, and P. L. Graumann, 2016. Super Resolution Fluorescence Microscopy and Tracking of Bacterial Flotillin (Reggie) Paralogs Provide Evidence for Defined-Sized Protein Microdomains within the Bacterial Membrane but Absence of Clusters Containing Detergent-Resistant Proteins. PLoS genetics 12:e1006116.

[104] de Pedro, M. A., C. G. Grünfelder, and H. Schwarz, 2004. Restricted Mobility of Cell Surface Proteins in the Polar Regions of Escherichia coli. Journal of bacteriology 186:2594–2602.

[105] Oddershede, L., J. K. Dreyer, S. Grego, S. Brown, and K. Berg-Sørensen, 2002. The motion of a single molecule, the lambda-receptor, in the bacterial outer membrane. Biophysj 83:3152–3161.

[106] Gibbs, K. A., D. D. Isaac, J. Xu, R. W. Hendrix, T. J. Silhavy, and J. A. Theriot, 2004. Complex spatial distribution and dynamics of an abundant Escherichia coli outer membrane protein, LamB. Molecular microbiology 53:1771–1783.

[107] Spector, J., S. Zakharov, Y. Lill, O. Sharma, W. A. Cramer, and K. Ritchie, 2010. Mobility of BtuB and OmpF in the Escherichia coli

outer membrane: implications for dynamic formation of a translocon complex. Biophysical journal 99:3880–3886.

[108] Rassam, P., N. A. Copeland, O. Birkholz, C. Tóth, M. Chavent, A. L. Duncan, S. J. Cross, N. G. Housden, R. Kaminska, U. Seger, D. M. Quinn, T. J. Garrod, M. S. P. Sansom, J. Piehler, C. G. Baumann, and C. Kleanthous, 2015. Supramolecular assemblies underpin turnover of outer membrane proteins in bacteria. Nature 523:333–336.

[109] Verhoeven, G. S., M. Dogterom, and T. den Blaauwen, 2013. Absence of long-range diffusion of OmpA in E. coli is not caused by its peptidoglycan binding domain. BMC microbiology 13:66.

[110] Brass, J. M., C. F. Higgins, M. Foley, P. A. Rugman, J. Birmingham, and P. B. Garland, 1986. Lateral diffusion of proteins in the periplasm of Escherichia coli. Journal of bacteriology 165:787–795.

[111] Zhang, L. C., Y. F. Chen, W. L. Chen, and C. C. Zhang, 2008. Existence of periplasmic barriers preventing green fluorescent protein diffusion from cell to cell in the cyanobacterium Anabaena sp. strain PCC 7120. Molecular microbiology 70:814–823.

[112] Zhang, L. C., V. Risoul, A. Latifi, J. M. Christie, and C. C. Zhang, 2013. Exploring the size limit of protein diffusion through the periplasm in cyanobacterium Anabaena sp. PCC 7120 using the 13 kDa iLOV fluorescent protein. Research in microbiology 164:710–717.

[113] Niu, L. and J. Yu, 2008. Investigating intracellular dynamics of FtsZ cytoskeleton with photoactivation single-molecule tracking. Biophysical journal 95:2009–2016.

[114] Wheeler, R. T. and L. Shapiro, 1999. Differential localization of two histidine kinases controlling bacterial cell differentiation. Molecular cell 4:683–694.

[115] Deich, J., E. M. Judd, H. H. McAdams, and W. E. Moerner, 2004. Visualization of the movement of single histidine kinase molecules in live Caulobacter cells. Proceedings of the National Academy of Sciences 101:15921–15926.

[116] Bolhuis, A., J. E. Mathers, J. D. Thomas, C. M. Barrett, and C. Robinson, 2001. TatB and TatC form a functional and structural unit of the

twin-arginine translocase from Escherichia coli. Journal of Biological Chemistry 276:20213–20219.

[117] Zhang, F., G. M. Lee, and K. Jacobson, 1993. Protein lateral mobility as a reflection of membrane microstructure. BioEssays : news and reviews in molecular, cellular and developmental biology 15:579–588.

[118] Leake, M. C., N. P. Greene, R. M. Godun, T. Granjon, G. Buchanan, S. Chen, R. M. Berry, T. Palmer, and B. C. Berks, 2008. Variable stoichiometry of the TatA component of the twin-arginine protein transport system observed by in vivo single-molecule imaging. Proceedings of the National Academy of Sciences of the United States of America 105:15376–15381.

[119] Oswald, F., A. Varadarajan, H. Lill, E. J. G. Peterman, and Y. J. M. Bollen, 2016. MreB-Dependent Organization of the E. coli Cytoplasmic Membrane Controls Membrane Protein Diffusion. Biophysical journal 110:1139–1149.

[120] Saffman, P. G. and M. Delbrück, 1975. Brownian motion in biological membranes. Proceedings of the National Academy of Sciences 72:3111–3113.

[121] Lucena, D., M. Mauri, F. Schmidt, B. Eckhardt, and P. L. Graumann, 2018. Microdomain formation is a general property of bacterial membrane proteins and induces heterogeneity of diffusion patterns :1–17.

[122] Lenn, T., M. C. Leake, and C. W. Mullineaux, 2008. Clustering and dynamics of cytochrome bd-I complexes in the Escherichia coli plasma membrane in vivo. Molecular microbiology 70:1397–1407.

[123] Oh, D., Y. Yu, H. Lee, B. L. Wanner, and K. Ritchie, 2014. Dynamics of the Serine Chemoreceptor in the Escherichia coli Inner Membrane: A High-Speed Single-Molecule Tracking Study. Biophysj 106:145–153.

[124] Chichili, G. R. and W. Rodgers, 2009. Cytoskeleton-membrane interactions in membrane raft structure. Cellular and molecular life sciences : CMLS 66:2319–2328.

[125] Goiko, M., J. R. de Bruyn, and B. Heit, 2016. Short-Lived Cages Restrict Protein Diffusion in the Plasma Membrane. Nature Publishing Group 6:34987.

[126] Sauer, M., 2013. Localization microscopy coming of age: from concepts to biological impact. J Cell Sci 126:3505–3513.

[127] Gahlmann, A. and W. E. Moerner, 2013. Exploring bacterial cell biology with single-molecule tracking and super-resolution imaging. Nature reviews. Microbiology 12:9–22.

[128] Izeddin, I., V. Récamier, L. Bosanac, I. I. Cissé, L. Boudarene, C. Dugast-Darzacq, F. Proux, O. Bénichou, R. Voituriez, O. Bensaude, M. Dahan, X. Darzacq, and R. H. Singer, 2014. Single-molecule tracking in live cells reveals distinct target-search strategies of transcription factors in the nucleus. eLife 3:e02230.

[129] Elf, J., G.-W. Li, and X. S. Xie, 2007. Probing Transcription Factor Dynamics at the Single-Molecule Level in a Living Cell. Science 316:1191–1194.

[130] Uphoff, S., R. Reyes-Lamothe, F. Garza de Leon, D. J. Sherratt, and A. N. Kapanidis, 2013. Single-molecule DNA repair in live bacteria. Proceedings of the National Academy of Sciences of the United States of America 110:8063–8068.

[131] Sanamrad, A., F. Persson, E. G. Lundius, D. Fange, A. H. Gynnå, and J. Elf, 2014. Single-particle tracking reveals that free ribosomal subunits are not excluded from the Escherichia coli nucleoid. Proceedings of the National Academy of Sciences of the United States of America 111:11413–11418.

[132] Plochowietz, A., I. Farrell, Z. Smilansky, B. S. Cooperman, and A. N. Kapanidis, 2016. In vivo single-RNA tracking shows that most tRNA diffuses freely in live bacteria. Nucleic acids research 45:gkw787–937.

[133] Liao, Y., J. W. Schroeder, B. Gao, L. A. Simmons, and J. S. Biteen, 2015. Single-molecule motions and interactions in live cells reveal target search dynamics in mismatch repair. Proceedings of the National Academy of Sciences of the United States of America 112:E6898–906.

[134] Bakshi, S., B. P. Bratton, and J. C. Weisshaar, 2011. Subdiffraction-limit study of Kaede diffusion and spatial distribution in live Escherichia coli. Biophysical journal 101:2535–2544.

[135] Chung, I., R. Akita, R. Vandlen, D. Toomre, J. Schlessinger, and I. Mellman, 2010. Spatial control of EGF receptor activation by reversible dimerization on living cells. Nature 464:783–787.

[136] Beausang, J. F., C. Zurla, C. Manzo, D. Dunlap, L. Finzi, and P. C. Nelson, 2007. DNA looping kinetics analyzed using diffusive hidden Markov model. Biophysj 92:L64–6.

[137] Endesfelder, U., K. Finan, S. J. Holden, P. R. Cook, A. N. Kapanidis, and M. Heilemann, 2013. Multiscale spatial organization of RNA polymerase in Escherichia coli. Biophysical journal 105:172–181.

[138] Jaqaman, K., D. Loerke, M. Mettlen, H. Kuwata, S. Grinstein, S. L. Schmid, and G. Danuser, 2008. Robust single-particle tracking in live-cell time-lapse sequences. Nature Methods 5:695–702.

[139] Fu, G., T. Huang, J. Buss, C. Coltharp, Z. Hensel, and J. Xiao, 2010. In vivo structure of the E. coli FtsZ-ring revealed by photoactivated localization microscopy (PALM). PLoS ONE 5:e12682.

[140] Szymborska, A., A. Szymborska, A. de Marco, A. de Marco, N. Daigle, N. Daigle, V. C. Cordes, V. C. Cordes, J. A. G. Briggs, J. A. G. Briggs, J. Ellenberg, and J. Ellenberg, 2013. Nuclear Pore Scaffold Structure Analyzed by Super-Resolution Microscopy and Particle Averaging. Science 341:655–658.

[141] Xu, K., G. Zhong, and X. Zhuang, 2013. Actin, spectrin, and associated proteins form a periodic cytoskeletal structure in axons .

[142] Xiao, J. and Y. F. Dufrêne, 2016. Optical and force nanoscopy in microbiology. Nature Microbiology 1:16186.

[143] Shroff, H., C. G. Galbraith, J. A. Galbraith, and E. Betzig, 2008. Live-cell photoactivated localization microscopy of nanoscale adhesion dynamics. Nature Methods 5:417–423.

[144] Coltharp, C., X. Yang, and J. Xiao, 2014. Quantitative analysis of single-molecule superresolution images. Current opinion in structural biology 28:112–121.

[145] Coles, B. C., S. Webb, and N. Schwartz, 2016. Characterisation of the effects of optical aberrations in single molecule techniques .

[146] von Diezmann, A., Y. Shechtman, and W. E. Moerner, 2017. Three-Dimensional Localization of Single Molecules for Super-Resolution Imaging and Single-Particle Tracking. Chemical Reviews 117:7244–7275.

[147] Shtengel, G., G. Shtengel, J. A. Galbraith, J. A. Galbraith, C. G. Galbraith, C. G. Galbraith, J. Lippincott-Schwartz, J. Lippincott-Schwartz, J. M. Gillette, J. M. Gillette, S. Manley, S. Manley, R. Sougrat, R. Sougrat, C. M. Waterman, C. M. Waterman, P. Kanchanawong, P. Kanchanawong, M. W. Davidson, M. W. Davidson, R. D. Fetter, R. D. Fetter, H. F. Hess, and H. F. Hess, 2009. Interferometric fluorescent super-resolution microscopy resolves 3D cellular ultrastructure. Proceedings of the National Academy of Sciences 106:3125–3130.

[148] Huang, B., B. Huang, W. Wang, W. Wang, M. Bates, M. Bates, X. Zhuang, and X. Zhuang, 2008. Three-Dimensional Super-Resolution Imaging by Stochastic Optical Reconstruction Microscopy. Science 319:810–813.

[149] Pavani, S. R. P., S. R. P. Pavani, M. A. Thompson, M. A. Thompson, J. S. Biteen, J. S. Biteen, S. J. Lord, S. J. Lord, N. Liu, N. Liu, R. J. Twieg, R. J. Twieg, R. Piestun, R. Piestun, and W. E. Moerner, 2009. Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function. Proceedings of the National Academy of Sciences 106:2995–2999.

[150] Ram, S., P. Prabhat, J. Chao, E. Sally Ward, E. S. Ward, and R. J. Ober, 2008. High Accuracy 3D Quantum Dot Tracking with Multifocal Plane Microscopy for the Study of Fast Intracellular Dynamics in Live Cells. Biophysical journal 95:6025–6043.

[151] Brown, T. A., T. A. Brown, A. N. Tkachuk, A. N. Tkachuk, G. Shtengel, G. Shtengel, B. G. Kopek, B. G. Kopek, D. F. Bogenhagen, D. F. Bogenhagen, H. F. Hess, H. F. Hess, D. A. Clayton, and D. A. Clayton, 2011. Superresolution Fluorescence Imaging of Mitochondrial Nucleoids Reveals Their Spatial Range, Limits, and Membrane Interaction. Molecular and Cellular Biology 31:4994–5010.

[152] Lyu, Z., C. Coltharp, X. Yang, and J. Xiao, 2016. Influence of FtsZ GTPase activity and concentration on nanoscale Z-ring structure in vivorevealed by three-dimensional Superresolution imaging. Biopolymers 105:725–734.

[153] Ovesný, M., P. Křížek, J. Borkovec, Z. Svindrych, and G. M. Hagen, 2014. ThunderSTORM: a comprehensive ImageJ plug-in for PALM and STORM data analysis and super-resolution imaging. Bioinformatics (Oxford, England) 30:2389–2390.

[154] Proppert, S., S. Wolter, T. Holm, and T. Klein, 2014. Cubic B-spline calibration for 3D super-resolution measurements using astigmatic imaging .

[155] Shaevitz, J. W., 2009. Bayesian estimation of the axial position in astigmatism-based three-dimensional particle tracking .

[156] Holden, S. J., K. M. Douglass, S. Manley, and L. Carlini, 2015. Correction of a Depth-Dependent Lateral Distortion in 3D Super-Resolution Imaging. PLoS ONE 10:e0142949.

[157] Liu, S., E. B. Kromann, W. D. Krueger, and J. Bewersdorf, 2013. Three dimensional single molecule localization using a phase retrieved pupil function .

[158] Small, A. and S. Stahlheber, 2014. Fluorophore localization algorithms for super-resolution microscopy. Nature Methods 11:267–279.

[159] Ober, R. J., S. Ram, and E. S. Ward, 2004. Localization Accuracy in Single-Molecule Microscopy. Biophysical journal 86:1185–1200.

[160] Besseling, T. H., J. Jose, A. V. BLAADEREN, and A. Van Blaaderen, 2014. Methods to calibrate and scale axial distances in confocal microscopy as a function of refractive index. Journal of Microscopy 257:142–150.

[161] Baddeley, D. and J. Bewersdorf, 2018. Biological Insight from Super-Resolution Microscopy: What We Can Learn from Localization-Based Images. Annual review of biochemistry 87:965–989.

[162] Sauer, M. and M. Heilemann, 2017. Single-Molecule Localization Microscopy in Eukaryotes. Chemical Reviews 117:7478–7509.

[163] Endesfelder, U., K. Finan, S. J. Holden, P. R. Cook, A. N. Kapanidis, and M. Heilemann, 2013. Multiscale Spatial Organization of RNA Polymerase in Escherichia coli. Biophysj 105:172–181.

[164] Chen, X., M. Wei, M. M. Zheng, J. Zhao, H. Hao, L. Chang, P. Xi, and Y. Sun, 2016. Study of RNA Polymerase II Clustering inside Live-Cell Nuclei Using Bayesian Nanoscopy. ACS Nano 10:2447–2454.

[165] Weng, X. and J. Xiao, 2014. Spatial organization of transcription in bacterial cells. Trends in genetics 30:287–297.

[166] Garcia-Parajo, M. F., A. Cambi, J. A. Torreno-Pina, N. Thompson, and K. Jacobson, 2014. Nanoclustering as a dominant feature of plasma membrane organization. J Cell Sci 127:4995–5005.

[167] Coltharp, C., J. Buss, T. M. Plumer, and J. Xiao, 2016. Defining the rate-limiting processes of bacterial cytokinesis. Proceedings of the National Academy of Sciences 113:E1044–E1053.

[168] Buss, J., C. Coltharp, T. Huang, C. Pohlmeyer, S.-C. Wang, C. Hatem, and J. Xiao, 2013. In vivoorganization of the FtsZ-ring by ZapA and ZapB revealed by quantitative super-resolution microscopy. Molecular microbiology 89:1099–1120.

[169] Buss, J., C. Coltharp, G. Shtengel, X. Yang, H. Hess, and J. Xiao, 2015. A multi-layered protein network stabilizes the Escherichia coli FtsZ-ring and modulates constriction dynamics. PLoS genetics 11:e1005128.

[170] Spühler, I. A., G. M. Conley, F. Scheffold, and S. G. Sprecher, 2016. Super Resolution Imaging of Genetically Labeled Synapses in Drosophila Brain Tissue. Frontiers in cellular neuroscience 10:142.

[171] Bar-On, D., S. Wolter, S. van de Linde, M. Heilemann, G. Nudelman, E. Nachliel, M. Gutman, M. Sauer, and U. Ashery, 2012. Super-resolution imaging reveals the internal architecture of nano-sized syntaxin clusters. Journal of Biological Chemistry 287:27158–27167.

[172] Xu, K., G. Zhong, and X. Zhuang, 2013. Actin, spectrin, and associated proteins form a periodic cytoskeletal structure in axons. Science 339:452–456.

[173] Xie, X., M. P. Cosma, and M. Lakadamyali, 2017. ScienceDirect Super resolution imaging of chromatin in pluripotency, differentiation, and reprogramming. Current opinion in genetics & development 46:186–193.

[174] Spahn, C., U. Endesfelder, and M. Heilemann, 2014. Super-resolution imaging of Escherichia coli nucleoids reveals highly structured and asymmetric segregation during fast growth. Journal of structural biology 185:243–249.

[175] Lehmann, M., S. Rocha, B. Mangeat, F. Blanchet, H. Uji-I, J. Hofkens, and V. Piguet, 2011. Quantitative multicolor super-resolution microscopy reveals tetherin HIV-1 interaction. PLoS pathogens 7:e1002456.

[176] Annibale, P., M. Scarselli, A. Kodiyan, and A. Radenovic, 2010. Photoactivatable Fluorescent Protein mEos2 Displays Repeated Photoactivation after a Long-Lived Dark State in the Red Photoconverted Form. The Journal of Physical Chemistry Letters 1:1506–1510.

[177] Baumgart, F., A. M. Arnold, K. Leskovar, K. Staszek, M. Fölser, J. Weghuber, H. Stockinger, and G. J. Schütz, 2016. Varying label density allows artifact-free analysis of membrane-protein nanoclusters. Nature Methods 13:661–664.

[178] Coltharp, C., R. P. Kessler, and J. Xiao, 2012. Accurate Construction of Photoactivated Localization Microscopy (PALM) Images for Quantitative Measurements. PLoS ONE 7:e51725–16.

[179] Sengupta, P., T. Jovanovic-Talisman, D. Skoko, M. Renz, S. L. Veatch, and J. Lippincott-Schwartz, 2011. Probing protein heterogeneity in the plasma membrane using PALM and pair correlation analysis. Nature Methods 8:969–975.

[180] Puchner, E. M., J. M. Walter, R. Kasper, B. Huang, and W. A. Lim, 2013. Counting molecules in single organelles with superresolution microscopy allows tracking of the endosome maturation trajectory. Proceedings of the National Academy of Sciences of the United States of America 110:16015–16020.

[181] Annibale, P., S. Vanni, M. Scarselli, U. Rothlisberger, and A. Radenovic, 2011. Identification of clustering artifacts in photoactivated localization microscopy. Nature Publishing Group 8:527–528.

[182] Hartwich, T. M. P., F. V. Subach, L. Cooley, V. V. Verkhusha, and J. Bewersdorf, 2013. Determination of two-photon photoactivation rates of fluorescent proteins. Physical chemistry chemical physics : PCCP 15:14868–14872.

[183] Rollins, G. C., J. Y. Shin, C. Bustamante, and S. Pressé, 2015. Stochastic approach to the molecular counting problem in superresolution microscopy. Proceedings of the National Academy of Sciences of the United States of America 112:E110–8.

[184] Hummer, G., F. Fricke, and M. Heilemann, 2016. Model-independent counting of molecules in single-molecule localization microscopy. Molecular biology of the cell 27:3637–3644.

[185] Nino, D., N. Rafiei, Y. Wang, A. Zilman, and J. N. Milstein, 2017. Molecular Counting with Localization Microscopy: A Bayesian Estimate Based on Fluorophore Statistics. Biophysj 112:1777–1785.

[186] Zhengxi Huang, Dongmei Ji, S. Wang, , A. Xia, Felix Koberling, M. Patting, , and R. Erdmann, 2005. Spectral Identification of Specific Photophysics of Cy5 by Means of Ensemble and Single Molecule Measurements. The Journal of Physical Chemistry A 110:45–50.

[187] Widengren, J., A. Chmyrov, C. Eggeling, P.-Å. Löfdahl, and C. A. M. Seidel, 2007. Strategies to Improve Photostabilities in Ultrasensitive Fluorescence Spectroscopy. The Journal of Physical Chemistry A 111:429–440.

[188] Vogelsang, J., R. Kasper, C. Steinhauer, B. Person, M. Heilemann, M. Sauer, and P. Tinnefeld, 2008. A Reducing and Oxidizing System Minimizes Photobleaching and Blinking of Fluorescent Dyes. Angewandte Chemie International Edition 47:5465–5469.

[189] Veatch, S. L., B. B. Machta, S. A. Shelby, E. N. Chiang, D. A. Holowka, and B. A. Baird, 2012. Correlation Functions Quantify Super-Resolution Images and Estimate Apparent Clustering Due to Over-Counting. PLoS ONE 7:e31457.

[190] Spahn, C., F. Herrmannsdörfer, T. Kuner, and M. Heilemann, 2016. Temporal accumulation analysis provides simplified artifact-free analysis of membrane-protein nanoclusters. Nature Methods 13:963–964.

[191] Mo, G. C. H., B. Ross, F. Hertel, P. Manna, X. Yang, E. Greenwald, C. Booth, A. M. Plummer, B. Tenner, Z. Chen, Y. Wang, E. J. Kennedy, P. A. Cole, K. G. Fleming, A. Palmer, R. Jimenez, J. Xiao, P. Dedecker, and J. Zhang, 2017. Genetically encoded biosensors for visualizing live-cell biochemical activity at super-resolution. Nature Methods 14:427–434.

[192] Zhang, J. and M. S. Shapiro, 2015. Mechanisms and dynamics of AKAP79/150-orchestrated multi-protein signalling complexes in brain and peripheral nerve. The Journal of Physiology 594:31–37.

[193] Zhang, J., C. M. Carver, F. S. Choveau, and M. S. Shapiro, 2016. Clustering and Functional Coupling of Diverse Ion Channels and Signaling Proteins Revealed by Super- resolution STORM Microscopy in Neurons. Neuron 92:461–478.

[194] Habuchi, S., R. Ando, P. Dedecker, W. Verheijen, H. Mizuno, A. Miyawaki, and J. Hofkens, 2005. Reversible single-molecule photoswitching in the GFP-like fluorescent protein Dronpa. Proceedings of the National Academy of Sciences 102:9511–9516.

[195] Malagon, F., 2013. RNase III is required for localization to the nucleoid of the 5' pre-rRNA leader and for optimal induction of rRNA synthesis in E. coli. RNA (New York, N.Y.) 19:1200–1207.

[196] Wooten, M., J. Snedeker, Z. F. Nizami, X. Yang, R. Ranjan, E. Urban, J. M. Kim, J. Gall, J. Xiao, and X. Chen, 2019. Asymmetric histone inheritance via strand-specific incorporation and biased replication fork movement. Nature Structural & Molecular Biology 26:732–743.

[197] Cella Zanacchi, F., C. Manzo, R. Magrassi, N. D. Derr, and M. Lakadamyali, 2019. Quantifying Protein Copy Number in Super Resolution Using an Imaging-Invariant Calibration. Biophysj 116:2195–2203.

[198] Datsenko, K. A. and B. L. Wanner, 2000. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proceedings of the National Academy of Sciences 97:6640–6645.

[199] Hensel, Z., X. Fang, and J. Xiao, 2013. Single-molecule Imaging of Gene Regulation In vivo Using Cotranslational Activation by Cleavage (CoTrAC). JoVE (Journal of Visualized Experiments) :e50042.

[200] Skinner, S. O., L. A. Sepúlveda, H. Xu, and I. Golding, 2013. Measuring mRNA copy number in individual Escherichia coli cells using single-molecule fluorescent in situ hybridization. Nature Protocols 8:1100–1113.

[201] Dempsey, G. T., J. C. Vaughan, K. H. Chen, M. Bates, and X. Zhuang, 2011. Evaluation of fluorophores for optimal performance in localization-based super-resolution imaging. Nature Methods 8:1027–1036.

[202] Malagon, F., 2013. RNase III is required for localization to the nucleoid of the 5' pre-rRNA leader and for optimal induction of rRNA synthesis in E. coli. RNA (New York, N.Y.) 19:1200–1207.

[203] Hensel, Z., X. Weng, A. C. Lagda, and J. Xiao, 2013. Transcription-Factor-Mediated DNA Looping Probed by High-Resolution, Single-Molecule Imaging in Live E. coli Cells. PLoS Biology 11:e1001591–17.

[204] Sage, D., H. Kirshner, T. Pengo, N. Stuurman, J. Min, S. Manley, and M. Unser, 2015. Quantitative evaluation of software packages for single-molecule localization microscopy. Nature Methods 12:717–724.

[205] Lyu, Z., C. Coltharp, X. Yang, and J. Xiao, 2016. Influence of FtsZ GTPase activity and concentration on nanoscale Z-ring structure in vivorevealed by three-dimensional Superresolution imaging. Biopolymers 105:725–734.

[206] Nahidiazar, L., A. V. Agronskaia, J. Broertjes, B. van den Broek, and K. Jalink, 2016. Optimizing Imaging Conditions for Demanding Multi-Color Super Resolution Localization Microscopy. PLoS ONE 11:e0158884.

[207] Schneider, C. A., W. S. Rasband, K. E. N. methods, and 2012. NIH Image to ImageJ: 25 years of image analysis. nature.com .

[208] Elowitz, M. B. and S. Leibler, 2000. A synthetic oscillatory network of transcriptional regulators. Nature 403:335–8.

[209] Elowitz, M. B., A. J. Levine, E. D. Siggia, and P. S. Swain, 2002. Stochastic gene expression in a single cell. Science 297:1183–1186.

[210] Yu, J., J. Xiao, X. Ren, K. Lao, and X. S. Xie, 2006. Probing gene expression in live cells, one protein molecule at a time. Science 311:1600–1603.

[211] Cai, L., N. Friedman, and X. S. Xie, 2006. Stochastic protein expression in individual cells at the single molecule level. Nature 440:358–62.

[212] Raj, A. and A. van Oudenaarden, 2008. Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. Cell 135:216–226.

[213] Sanchez, A. and I. Golding, 2013. Genetic determinants and cellular constraints in noisy gene expression. Science 342:1188–93.

[214] Taniguchi, Y., P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie, 2010. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science 329:533–538.

[215] Swain, P. S., M. B. Elowitz, and E. D. Siggia, 2002. Intrinsic and extrinsic contributions to stochasticity in gene expression. Proceedings of the National Academy of Sciences 99:12795–12800.

[216] Hilfinger, A. and J. Paulsson, 2011. Separating intrinsic from extrinsic fluctuations in dynamic biological systems. Proc Natl Acad Sci USA 108:12167–72.

[217] Shahrezaei, V. and P. S. Swain, 2008. Analytical distributions for stochastic gene expression. Proceedings of the National Academy of Sciences 105:17256–17261.

[218] Singh, A., B. Razooky, C. D. Cox, M. L. Simpson, and L. S. Weinberger, 2010. Transcriptional Bursting from the HIV-1 Promoter Is a Significant Source of Stochastic Noise in HIV-1 Gene Expression. Biophysical journal 98:L32–L34.

[219] Hensel, Z., H. Feng, B. Han, C. Hatem, J. Wang, and J. Xiao, 2012. Stochastic expression dynamics of a transcription factor revealed by single-molecule noise analysis. Nature Structural & Molecular Biology 19:797–802.

[220] Assaf, M., E. Roberts, Z. Luthey-Schulten, and N. Goldenfeld, 2013. Extrinsic noise driven phenotype switching in a self-regulating gene. Phys Rev Lett 111:058102.

[221] Jones, D. L., R. C. Brewster, and R. Phillips, 2014. Promoter architecture dictates cell-to-cell variability in gene expression. Science 346:1533–1536.

[222] Munsky, B., G. Neuert, and A. van Oudenaarden, 2012. Using gene expression noise to understand gene regulation. Science 336:183–7.

[223] Golding, I., J. Paulsson, S. M. Zawilski, and E. C. Cox, 2005. Real-Time Kinetics of Gene Activity in Individual Bacteria. Cell 123:1025–1036.

[224] Raj, A., C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi, 2006. Stochastic mRNA synthesis in mammalian cells. PLoS Biol 4:e309.

[225] So, L.-h., A. Ghosh, C. Zong, L. A. Sepúlveda, R. Segev, and I. Golding, 2011. General properties of transcriptional time series in Escherichia coli. Nature Genetics 43:554–560.

[226] Assaf, M., E. Roberts, and Z. Luthey-Schulten, 2011. Determining the Stability of Genetic Switches: Explicitly Accounting for mRNA Noise. Physical Review Letters 106:248102.

[227] Liu, L. F. and J. C. Wang, 1987. Supercoiling of the DNA template during transcription. Proceedings of the National Academy of Sciences 84:7024–7027.

[228] McAdams, H. H. and A. Arkin, 1997. Stochastic mechanisms in gene expression. Proc Natl Acad Sci USA 94:814–9.

[229] Paulsson, J., 2005. Models of stochastic gene expression. Phys Life Rev 2:157–75.

[230] Friedman, N., L. Cai, and X. Xie, 2006. Linking Stochastic Dynamics to Population Distribution: An Analytical Framework of Gene Expression. Physical Review Letters 97:168302.

[231] Roberts, E., A. Magis, J. O. Ortiz, W. Baumeister, and Z. Luthey-Schulten, 2011. Noise contributions in an inducible genetic switch: a whole-cell simulation study. PLoS Comput Biol 7:e1002010.

[232] Gong, P., E. A. Esposito, and C. T. Martin, 2004. Initial bubble collapse plays a key role in the transition to elongation in T7 RNA polymerase. Journal of Biological Chemistry 279:44277–44285.

[233] Bandwar, R. P. and S. S. Patel, 2002. The Energetics of Consensus Promoter Opening by T7 RNA Polymerase. Journal of molecular biology 324:63–72.

[234] Djordjevic, M. and R. Bundschuh, 2008. Formation of the Open Complex by Bacterial RNA Polymerase—A Quantitative Model. Biophysical journal 94:4233–4248.

[235] Record, M. T., W. S. Reznikoff, and M. L. Craig, 1996. Escherichia coli RNA polymerase (Es70), promoters, and the kinetics of the steps of transcription initiation. Cellular and molecular biology 1:792.

[236] Shepherd, N., P. Dennis, and H. Bremer, 2001. Cytoplasmic RNA Polymerase in Escherichia coli. Journal of bacteriology 183:2527–2534.

[237] Ujvari, A. and C. T. Martin, 1996. Thermodynamic and Kinetic Measurements of Promoter Binding by T7 RNA Polymerase†. Biochemistry 35:14574–14582.

[238] Gagua, A. V., B. N. Belintsev, and Y. L. Lyubchenko, 1981. Effect of base-pair stability on the melting of superhelical DNA. Nature 294:662–663.

[239] Ramirez-Tapia, L. E. and C. T. Martin, 2012. New insights into the mechanism of initial transcription: the T7 RNA polymerase mutant P266L transitions to elongation at longer RNA lengths than wild type. Journal of Biological Chemistry 287:37352–37361.

[240] Sen, S. and R. Majumdar, 1988. Statistical mechanical theory of melting transition in supercoiled DNA. Biopolymers 27:1479–1489.

[241] Sen, S., A. Lahiri, and R. Majumdar, 1992. Melting characteristics of highly supercoiled DNA. Biophysical chemistry 42:229–234.

[242] Benham, C. J., 1977. Elastic model of supercoiling. Proceedings of the National Academy of Sciences 74:2397–2401.

[243] Benham, C. J., 1979. Torsional stress and local denaturation in supercoiled DNA. Proceedings of the National Academy of Sciences 76:3870–3874.

[244] Benham, C. J., 1980. Kinetics of reactions involving DNA containing stress-induced single-stranded regions. Biopolymers 19:2143–2164.

[245] Depew, D. E. and J. C. Wang, 1975. Conformational fluctuations of DNA helix. Proceedings of the National Academy of Sciences 72:4275–4279.

[246] Tsao, Y.-P., H.-Y. Wu, and L. F. Liu, 1989. Transcription-driven supercoiling of DNA: Direct biochemical evidence from in vitro studies. Cell 56:111–118.

[247] Roberts, E., A. Magis, J. O. Ortiz, W. Baumeister, and Z. Luthey-Schulten, 2011. Noise Contributions in an Inducible Genetic Switch: A Whole-Cell Simulation Study. PLoS Computational Biology 7:e1002010.

[248] Stamatakis, M. and N. V. Mantzaris, 2009. Comparison of Deterministic and Stochastic Models of the lac Operon Genetic Network. Biophysical journal 96:887–906.

[249] Paulsson, J., 2005. Models of stochastic gene expression. Physics of life reviews 2:157–175.

[250] Cheng, B., C.-X. Zhu, C. Ji, A. Ahumada, and Y.-C. Tse-Dinh, 2003. Direct interaction between Escherichia coli RNA polymerase and the zinc ribbon domains of DNA topoisomerase I. Journal of Biological Chemistry 278:30705–30710.

[251] Gillespie, D. T., 1977. Exact stochastic simulation of coupled chemical reactions. The journal of physical chemistry 81:2340–2361.

[252] Roberts, E., J. E. Stone, and Z. Luthey-Schulten, 2013. Lattice microbes: High-performance stochastic simulation method for the reaction-diffusion master equation. Journal of Computational Chemistry 34:245–255.

[253] Higgins, N. P. and N. R. Cozzarelli, 1982. The binding of gyrase to DNA: analysis by retention by nitrocellulose filters. Nucleic acids research 10:6833–6847.

[254] Maxwell, A. and M. Gellert, 1984. The DNA dependence of the ATPase activity of DNA gyrase. Journal of Biological Chemistry 259:14472–14480.

[255] Sengupta, S. and V. Nagaraja, 2008. YacG from Escherichia coli is a specific endogenous inhibitor of DNA gyrase. Nucleic acids research 36:4310–4316.

[256] Nakanishi, A., T. Oshida, T. Matsushita, S. Imajoh-Ohmi, and T. Ohnuki, 1998. Identification of DNA Gyrase Inhibitor (GyrI) in-Escherichia coli. Journal of Biological Chemistry 273:1933–1938.

[257] Hardy, C. D. and N. R. Cozzarelli, 2005. A genetic selection for supercoiling mutants of Escherichia coli reveals proteins implicated in chromosome structure. Molecular microbiology 57:1636–1652.

[258] Fisher, L. M., K. Mizuuchi, M. H. O'Dea, H. Ohmori, and M. Gellert, 1981. Site-specific interaction of DNA gyrase with DNA. Proceedings of the National Academy of Sciences 78:4165–4169.

[259] Jeong, K. S., J. Ahn, and A. B. Khodursky, 2004. Spatial patterns of transcriptional activity in the chromosome of Escherichia coli. Genome Biol 5:R86.

[260] Bryant, J. A., L. E. Sellars, S. J. W. Busby, and D. J. Lee, 2014. Chromosome position effects on gene expression in Escherichia coli K-12. Nucleic Acids Res 42:11383–92.

[261] Iber, D., 2006. A quantitative study of the benefits of co-regulation using the spoIIA operon as an example. Molecular Systems Biology 2:43.

[262] Liang, L. W., R. Hussein, D. H. S. Block, and H. N. Lim, 2013. Minimal effect of gene clustering on expression in Escherichia coli. Genetics 193:453–465.

[263] Singh, A., 2011. Negative Feedback Through mRNA Provides the Best Control of Gene-Expression Noise. NanoBioscience, IEEE Transactions on 10:194–200.

[264] Thattai, M. and A. van Oudenaarden, 2001. Intrinsic noise in gene regulatory networks. Proceedings of the National Academy of Sciences 98:8614–8619.

[265] Bickmore, W. A., 2013. The Spatial Organization of the Human Genome. Annual Review of Genomics and Human Genetics 14:67–84.

[266] Bonev, B., N. Mendelson Cohen, Q. Szabo, L. Fritsch, G. L. Papadopoulos, Y. Lubling, X. Xu, X. Lv, J.-P. Hugnot, A. Tanay, and G. Cavalli, 2017. Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell 171:557–572.e24.

[267] Rowley, M. J., M. H. Nichols, X. Lyu, M. A.-K. M. cell, and 2017. Evolutionarily conserved principles predict 3D chromatin organization. cell.com .

[268] Rowley, M. J. and V. G. Corces, 2018. Organizational principles of 3D genome architecture. Nature Reviews Genetics 19:789–800.

[269] Davidson, I. F., B. Bauer, D. Goetz, W. Tang, G. Wutz, and J.-M. Peters, 2019. DNA loop extrusion by human cohesin. Science 366:1338–1345.

[270] Le, T. B. K., M. V. Imakaev, L. A. Mirny, and M. T. Laub, 2013. High-Resolution Mapping of the Spatial Organization of a Bacterial Chromosome. Science 342:731–734.

[271] Kim, S., B. Beltran, I. Irnov, and C. Jacobs-Wagner, 2019. Long-Distance Cooperative and Antagonistic RNA Polymerase Dynamics via DNA Supercoiling. Cell 179:106–119.e16.

[272] Dandanell, G., P. Valentin-Hansen, J. E. Larsen, and K. Hammer, 1987. Long-range cooperativity between gene regulatory sequences in a prokaryote. Nature 325:823–826.

[273] Dunn, T. M., S. Hahn, S. Ogden, and R. F. Schleif, 1984. An operator at -280 base pairs that is required for repression of araBAD operon promoter: addition of DNA helical turns between the operator and promoter cyclically hinders repression. Proceedings of the National Academy of Sciences 81:5017–5020.

[274] Hill, B. M., 1998. The function of auxiliary operators. Molecular microbiology 29:13–18.

[275] Hao, N., K. E. Shearwin, and I. B. Dodd, 2019. Positive and Negative Control of Enhancer-Promoter Interactions by Other DNA Loops Generates Specificity and Tunability. Cell reports 26:2419–2433.e3.

[276] Révet, B., B. von Wilcken-Bergmann, H. Bessert, A. Barker, and B. Müller-Hill, 1999. Four dimers of λ repressor bound to two suitably spaced pairs of λ operators form octamers and DNA loops over large distances. Current biology : CB 9:151–154.

[277] Hao, N., K. Sneppen, K. S. N. acids, and 2017. Efficient chromosomal-scale DNA looping in Escherichia coli using multiple DNA-looping elements. academic.oup.com .

[278] Kuhlman, T. E. and E. C. Cox, 2010. Site-specific chromosomal integration of large synthetic constructs. Nucleic acids research 38:e92–e92.

[279] Shahrezaei, V. and P. S. Swain, 2008. Analytical distributions for stochastic gene expression. Proceedings of the National Academy of Sciences of the United States of America 105:17256–17261.

[280] Rodriguez, J., G. Ren, C. R. Day, K. Zhao, C. C. Chow, and D. R. Larson, 2019. Intrinsic Dynamics of a Human Gene Reveal the Basis of Expression Heterogeneity. Cell 176:213–226.e18.

[281] Jones, D. and J. Elf, 2018. Bursting onto the scene? Exploring stochastic mRNA production in bacteria. Current opinion in microbiology 45:124–130.

[282] Ding, Y., C. Manzo, G. Fulcrand, F. Leng, D. Dunlap, and L. Finzi, 2014. DNA supercoiling: A regulatory signal for the λ repressor. Proceedings of the National Academy of Sciences 111:15402–15407.

[283] Yan, Y., Y. Ding, F. Leng, D. Dunlap, and L. Finzi, 2018. Protein-mediated loops in supercoiled DNA create large topological domains. Nucleic acids research 46:4417–4424.

[284] Fang, X., Q. Liu, C. Bohrer, Z. Hensel, W. Han, J. Wang, and J. Xiao, 2017. New Cell Fate Potentials and Switching Kinetics Uncovered in a Classic Bistable Genetic Switch. bioRxiv :1–23.

[285] Kuhlman, T. E. and E. C. Cox, 2010. Site-specific chromosomal integration of large synthetic constructs. Nucleic acids research 38:e92.

[286] Shapiro, L., H. H. McAdams, and R. Losick, 2009. Why and how bacteria localize proteins. Science 326:1225–1228.

[287] Losick, R. and L. Shapiro, 1999. Changing views on the nature of the bacterial cell: from biochemistry to cytology. 181:4143–4145.

[288] Bi, E. F. and J. Lutkenhaus, 1991. FtsZ ring structure associated with division in Escherichia coli. Nature 354:161–164.

[289] Cabrera, J. E. and D. J. Jin, 2003. The distribution of RNA polymerase in Escherichia coli is dynamic and sensitive to environmental cues. Molecular Microbiology 50:1493–1505.

[290] Lewis, P. J., S. D. Thaker, and J. Errington, 2000. Compartmentalization of transcription and translation in Bacillus subtilis. The EMBO journal 19:710–718.

[291] Bremer, H., P. Dennis, and M. Ehrenberg, 2003. Free RNA polymerase and modeling global transcription in Escherichia coli. Biochimie 85:597–609.

[292] Jin, D. J., C. Mata Martin, Z. Sun, C. Cagliero, and Y. N. Zhou, 2016. Nucleolus-like compartmentalization of the transcription machinery in fast-growing bacterial cells. Critical Reviews in Biochemistry and Molecular Biology 52:96–106.

[293] Cabrera, J. E. and D. J. Jin, 2006. Active transcription of rRNA operons is a driving force for the distribution of RNA polymerase in bacteria:

effect of extrachromosomal copies of rrnB on the in vivo localization of RNA polymerase. 188:4007–4014.

[294] Cook, P. R., 2010. A model for all genomes: the role of transcription factories. Journal of Molecular Biology 395:1–10.

[295] Marenduzzo, D., I. Faro-Trindade, and P. Cook, 2007. What are the molecular ties that maintain genomic loops? TRENDS in Genetics 23:126–133.

[296] Guenther, M. G., S. S. Levine, L. A. Boyer, R. Jaenisch, and R. A. Young, 2007. A Chromatin Landmark and Transcription Initiation at Most Promoters in Human Cells. Cell 130:77–88.

[297] Zeitlinger, J., A. Stark, M. Kellis, J.-W. Hong, S. Nechaev, K. Adelman, M. Levine, and R. A. Young, 2007. RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. Nature Genetics 39:1512–1516.

[298] Radonjic, M., J.-C. Andrau, P. Lijnzaad, P. Kemmeren, T. T. J. P. Kockelkorn, D. van Leenen, N. L. van Berkum, and F. C. P. Holstege, 2005. Genome-Wide Analyses Reveal RNA Polymerase II Located Upstream of Genes Poised for Rapid Response upon S. cerevisiae Stationary Phase Exit. Molecular cell 18:171–183.

[299] Muse, G. W., D. A. Gilchrist, S. Nechaev, R. Shah, J. S. Parker, S. F. Grissom, J. Zeitlinger, and K. Adelman, 2007. RNA polymerase is poised for activation across the genome. Nature Genetics 39:1507–1511.

[300] Mooney, R. A., S. E. Davis, J. M. Peters, J. L. Rowland, A. Z. Ansari, and R. Landick, 2009. Regulator trafficking on bacterial transcription units in vivo. Molecular cell 33:97–108.

[301] Reppas, N. B., J. T. Wade, G. M. Church, and K. Struhl, 2006. The transition between transcriptional initiation and elongation in E. coli is highly variable and often rate limiting. Molecular cell 24:747–757.

[302] Peano, C., J. Wolf, J. Demol, E. Rossi, L. Petiti, G. De Bellis, J. Geiselmann, T. Egli, S. Lacour, and P. Landini, 2015. Characterization of the Escherichia coli σS core regulon by Chromatin Immunoprecipitation-sequencing (ChIP-seq) analysis. Scientific reports :1–15.

[303] Shavkunov, K. S., I. S. Masulis, M. N. Tutukina, A. A. Deev, and O. N. Ozoline, 2009. Gains and unexpected lessons from genome-scale promoter mapping. Nucleic acids research 37:4919–4931.

[304] Huerta, A. M., M. P. Francino, E. Morett, and J. Collado-Vides, 2006. Selection for Unequal Densities of σ70 Promoter-Like Signals in Different Regions of Large Bacterial Genomes. PLoS genetics 2:e185–11.

[305] Huerta, A. M. and J. Collado-Vides, 2003. Sigma70 Promoters in Escherichia coli: Specific Transcription in Dense Regions of Overlapping Promoter-like Signals. Journal of Molecular Biology 333:261–278.

[306] Campbell, E. A., N. Korzheva, A. Mustaev, K. Murakami, S. Nair, A. Goldfarb, and S. A. Darst, 2001. Structural mechanism for rifampicin inhibition of bacterial rna polymerase. Cell 104:901–912.

[307] Cabrera, J. E., C. Cagliero, S. Quan, C. L. Squires, and D. J. Jin, 2009. Active transcription of rRNA operons condenses the nucleoid in Escherichia coli: examining the effect of transcription on nucleoid structure in the absence of transertion. 191:4180–4185.

[308] Le, T. B. K., M. V. Imakaev, L. A. Mirny, and M. T. Laub, 2013. High-resolution mapping of the spatial organization of a bacterial chromosome. Science 342:731–734.

[309] Stracy, M. and A. N. Kapanidis, 2017. Single-molecule and super-resolution imaging of transcription in living bacteria. Methods :1–12.

[310] Endesfelder, U., S. Malkusch, F. Fricke, and M. Heilemann, 2014. A simple method to estimate the average localization precision of a single-molecule localization microscopy experiment. Histochemistry and Cell Biology 141:629–638.

[311] Coltharp, C., R. P. Kessler, and J. Xiao, 2012. Accurate Construction of Photoactivated Localization Microscopy (PALM) Images for Quantitative Measurements. PLoS ONE 7:e51725.

[312] Wang, S., J. R. Moffitt, G. T. Dempsey, X. S. Xie, and X. Zhuang, 2014. Characterization and development of photoactivatable fluorescent proteins for single-molecule-based superresolution imaging. Proceedings of the National Academy of Sciences 111:8452–8457.

[313] Bohrer, C. H., X. Weng, K. Bettridge, Z. Lyu, R. McQuillen, X. Yang, and J. Xiao, 2018. A model-independent algorithm to extract blinking-free super-resolution images. (in submission) .

[314] Iwakura, Y., K. Ito, and A. Ishihama, 1974. Biosynthesis of RNA polymerase in Escherichia coli. I. Control of RNA polymerase content at various growth rates. Molecular & general genetics : MGG 133:1–23.

[315] Ishihama, A., 2000. Functional modulation of Escherichia coli RNA polymerase. Annual Review of Microbiology 54:499–518.

[316] Murakami, K. S., S. Masuda, E. A. Campbell, O. Muzzin, and S. A. Darst, 2002. Structural Basis of Transcription Initiation: An RNA Polymerase Holoenzyme-DNA Complex. Science 296:1285–1290.

[317] Jin, D. and J. Cabrera, 2006. Coupling the distribution of RNA polymerase to global gene regulation and the dynamic structure of the bacterial nucleoid in Escherichia coli. Journal of structural biology 156:284–291.

[318] Durfee, T., A.-M. Hansen, H. Zhi, F. R. Blattner, and D. J. Jin, 2008. Transcription profiling of the stringent response in Escherichia coli. 190:1084–1096.

[319] Hauryliuk, V., G. C. Atkinson, K. S. Murakami, T. Tenson, and K. Gerdes, 2015. Recent functional insights into the role of (p)ppGpp in bacterial physiology. Nature Publishing Group 13:298–309.

[320] Paul, B. J., W. Ross, T. Gaal, and R. L. Gourse, 2004. rRNA Transcription in Escherichia coli. Annual Review of Genetics 38:749–770.

[321] Gaal, T., B. P. Bratton, P. Sanchez-Vazquez, A. Sliwicki, K. Sliwicki, A. Vegel, R. Pannu, and R. L. Gourse, 2016. Colocalization of distant chromosomal loci in space in E. coli: a bacterial nucleolus. Genes & Development 30:2272–2285.

[322] Quan, S., O. Skovgaard, R. E. McLaughlin, E. T. Buurman, and C. L. Squires, 2015. Markerless Escherichia coli rrn Deletion Strains for Genetic Determination of Ribosomal Binding Sites. G3: Genes—Genomes—Genetics :1–13.

[323] Zaporojets, D., S. French, and C. L. Squires, 2003. Products Transcribed from Rearranged rrn Genes of Escherichia coli Can Assemble To Form Functional Ribosomes. Journal of Bacteriology 185:6921–6927.

[324] Asai, T., D. Zaporojets, C. Squires, and C. L. Squires, 1999. An Escherichia coli strain with all chromosomal rRNA operons inactivated: Complete exchange of rRNA genes between bacteria. Proceedings of the National Academy of Sciences of the United States of America 96:1971–1976.

[325] Severinova, E. e. a., 1998. Inhibition of Escherichia coli RNA Polymerase by Bacteriophage T4 AsiA :1–10.

[326] Gama-Castro, S., H. Salgado, A. Santos-Zavaleta, D. Ledezma-Tejeida, L. Muñiz-Rascado, J. S. García-Sotelo, K. Alquicira-Hernández, I. Martínez-Flores, L. Pannier, J. A. Castro-Mondragón, A. Medina-Rivera, H. Solano-Lira, C. Bonavides-Martínez, E. Pérez-Rueda, S. Alquicira-Hernández, L. Porrón-Sotelo, A. López-Fuentes, A. Hernández-Koutoucheva, V. D. Moral-Chávez, F. Rinaldi, and J. Collado-Vides, 2016. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic Acids Research 44:D133–D143.

[327] Sharma, U. K. and D. Chatterji, 2008. Differential mechanisms of binding of anti-sigma factors Escherichia coli Rsd and bacteriophage T4 AsiA to E. coli RNA polymerase lead to diverse physiological consequences. Journal of Bacteriology 190:3434–3443.

[328] Chai, Q., B. Singh, K. Peisker, N. Metzendorf, X. Ge, S. Dasgupta, and S. Sanyal, 2014. Organization of Ribosomes and Nucleoids in Escherichia coliCells during Growth and in Quiescence. The Journal of biological chemistry 289:11342–11352.

[329] Luijsterburg, M. S., M. F. White, R. van Driel, and R. T. Dame, 2008. The major architects of chromatin: architectural proteins in bacteria, archaea and eukaryotes. Critical reviews in biochemistry and molecular biology 43:393–418.

[330] Wang, X., P. M. Llopis, and D. Z. Rudner, 2013. Organization and segregation of bacterial chromosomes. Nature reviews. Genetics 14:191–203.

[331] Dame, R. T., O. J. Kalmykowa, and D. C. Grainger, 2011. Chromosomal macrodomains and associated proteins: implications for DNA organization and replication in gram negative bacteria. PLoS genetics 7:e1002123.

[332] Nollmann, M., N. J. Crisona, and P. B. Arimondo, 2007. Thirty years of Escherichia coli DNA gyrase: from in vivo function to single-molecule mechanism. Biochimie 89:490–499.

[333] Alt, S., L. A. Mitchenall, A. Maxwell, and L. Heide, 2011. Inhibition of DNA gyrase and DNA topoisomerase IV of Staphylococcus aureus and Escherichia coli by aminocoumarin antibiotics. Journal of Antimicrobial Chemotherapy 66:2061–2069.

[334] Collin, F., S. Karkare, and A. Maxwell, 2011. Exploiting bacterial DNA gyrase as a drug target: current state and perspectives. Applied microbiology and biotechnology 92:479–497.

[335] Wahle, E., K. Mueller, and E. ORR, 1985. Effect of Dna Gyrase Inactivation on Rna-Synthesis in Escherichia-Coli 162:458–460.

[336] Wahle, E. and K. Mueller, 1980. Involvement of DNA gyrase in rRNA synthesis in vivo. Molecular & general genetics : MGG 179:661–667.

[337] OOSTRA, B. A., A. J. VANVLIET, G. AB, and M. GRUBER, 1981. Enhancement of Ribosomal Ribonucleic-Acid Synthesis by Deoxyribonucleic-Acid Gyrase Activity in Escherichia-Coli 148:782–787.

[338] OOSTRA, B. A., G. AB, and M. GRUBER, 1980. Involvement of Dna Gyrase in the Transcription of Ribosomal-Rna. Nucleic Acids Research 8:4235–4246.

[339] Grigorova, I. L., N. J. Phleger, V. K. Mutalik, and C. A. Gross, 2006. Insights into transcriptional regulation and sigma competition from an equilibrium model of RNA polymerase binding to DNA. Proceedings of the National Academy of Sciences of the United States of America 103:5332–5337.

[340] Lau, I. F., S. R. Filipe, B. Søballe, O.-A. Økstad, F.-X. Barre, and D. J. Sherratt, 2004. Spatial and temporal organization of replicating Escherichia coli chromosomes. Molecular Microbiology 49:731–743.

[341] Stuger, R., C. L. Woldringh, C. C. van der Weijden, N. O. E. Vischer, B. M. Bakker, R. J. M. van Spanning, J. L. Snoep, and H. V. Westerhoff, 2002. DNA supercoiling by gyrase is linked to nucleoid compaction. Molecular biology reports 29:79–82.

[342] Peter, B. J., J. Arsuaga, A. M. Breier, A. B. Khodursky, P. O. Brown, and N. R. Cozzarelli, 2004. Genomic transcriptional response to loss of chromosomal supercoiling in Escherichia coli. Genome Biol 5:R87.

[343] SANZEY, B., 1979. Modulation of Gene-Expression by Drugs Affecting Deoxyribonucleic-Acid Gyrase 138:40–47.

[344] Feric, M., N. Vaidya, T. S. Harmon, D. M. Mitrea, L. Zhu, T. M. Richardson, R. W. Kriwacki, R. V. Pappu, and C. P. Brangwynne, 2016. Coexisting Liquid Phases Underlie Nucleolar Subcompartments. Cell 165:1686–1697.

[345] Berry, J., S. C. Weber, N. Vaidya, M. Haataja, and C. P. Brangwynne, 2015. RNA transcription modulates phase transition-driven nuclear body assembly. Proceedings of the National Academy of Sciences of the United States of America 112:E5237–E5245.

[346] Chen, J., K. M. Wassarman, S. Feng, K. Leon, A. Feklistov, J. T. Winkelman, Z. Li, T. Walz, E. A. Campbell, and S. A. Darst, 2017. 6S RNA Mimics B-Form DNA to Regulate Escherichia coli RNA Polymerase. Molecular cell 68:388–397.e6.

[347] Cavanagh, A. T. and K. M. Wassarman, 2014. 6S RNA, a Global Regulator of Transcription in Escherichia coli, Bacillus subtilis, and Beyond. Annual Review of Microbiology 68:45–60.

[348] Gill, S. C., S. E. Weitzel, and P. H. von Hippel, 1991. Escherichia coli sigma 70 and NusA proteins. I. Binding interactions with core RNA polymerase in solution and within the transcription complex. Journal of Molecular Biology 220:307–324.

[349] Vogel, U. and K. F. Jensen, 1997. NusA is required for ribosomal antitermination and for modulation of the transcription elongation rate of both antiterminated RNA and mRNA. The Journal of biological chemistry 272:12265–12271.

[350] Greive, S. J., A. F. Lins, and P. H. Von Hippel, 2005. Assembly of an RNA-Protein Complex BINDING OF NusB AND NusE (S10) PROTEINS TO boxA RNA NUCLEATES THE FORMATION OF THE ANTITERMINATION COMPLEX INVOLVED IN CONTROLLING rRNA TRANSCRIPTION IN ESCHERICHIA COLI. The Journal of biological chemistry 280:36397–36408.

[351] Stagno, J. R., A. S. Altieri, M. Bubunenko, S. G. Tarasov, J. Li, D. L. Court, R. A. Byrd, and X. Ji, 2011. Structural basis for RNA recognition by NusB and NusE in the initiation of transcription antitermination. Nucleic Acids Research 39:7803–7815.

[352] Drögemüller, J., M. Strauß, K. Schweimer, M. Jurk, P. Rösch, and S. H. Knauer, 2015. Determination of RNA polymerase binding surfaces of transcription factors by NMR spectroscopy. Scientific reports :1–14.

[353] Werner, F., 2012. A Nexus for Gene Expression—Molecular Mechanisms of Spt5 and NusG in the Three Domains of Life. Journal of Molecular Biology 417:13–27.

[354] Singh, N., M. Bubunenko, C. Smith, D. M. Abbott, A. M. Stringer, R. Shi, D. L. Court, and J. T. Wade, 2016. SuhB Associates with Nus Factors To Facilitate 30S Ribosome Biogenesis in Escherichia coli. mBio 7:e00114–16–9.

[355] Datsenko, K. A. and B. L. Wanner, 2000. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proceedings of the National Academy of Sciences 97:6640–6645.

[356] Hensel, Z., X. Weng, A. C. Lagda, and J. Xiao, 2013. Transcription-Factor-Mediated DNA Looping Probed by High-Resolution, Single-Molecule Imaging in Live E. coli Cells. PLoS Biology 11:e1001591.

[357] Hensel, Z., X. Fang, and J. Xiao, 2013. Single-molecule imaging of gene regulation in vivo using cotranslational activation by cleavage (CoTrAC). Journal of Visual Experiments :1–1.

[358] Xiao, J., J. Elf, G.-W. Li, J. Yu, and X. S. Xie, 2008. Imaging gene expression in living cells at the single-molecule level. In P. R. Selvin and T. Ha, editors, Single-Molecule Techniques A Laboratory Manual, Springer US, Boston, MA, 149–170.

[359] Xiao, J., 2009. Single-Molecule Imaging in Live Cells. In P. Hinterdorfer and A. Oijen, editors, Handbook of Single-Molecule Biophysics, Springer US, New York, NY, 43–93.

[360] Buss, J., C. Coltharp, and J. Xiao, 2013. Super-resolution Imaging of the Bacterial Division Machinery. Journal of Visual Experiments .

[361] Weng, X., 2018. place holder temporary :1–1.

Christopher H. Bohrer

cbohrer1@jhu.edu • +216-375-7805

Education:

2020 Ph.D, Biophysics and Biophysical Chemistry, Johns Hopkins, Baltimore, Maryland
2013 B.S Physics (magna cum laude), Physics, Kent State University, Kent, Ohio
2013 B.S Education (magna cum laude), Education, Kent State University, Kent, Ohio
2013 Minor of Biology, Biology, Kent State University, Kent, Ohio

Research Experience:

1. Super-Resolution Imaging/Method Development — JHU Dr. Xiao (2014-present) Developed/Used novel methodologies to study the influence of chromosome structure on transcription with super-resolution microscopy.
2. Computational Biophysics/Bioinformatics — JHU Dr. Roberts (2014-present) Developed/Used theory and computation to investigate the effects of local DNA structure on information propagation in bacteria.
3. Theoretical Molecular Biophysics — KSU Dr. Portman (2010-2013) Studied structural stability of various minor species of G-Quadruplex using various computational techniques.
4. Single-Molecule Biophysics — KSU Dr. Mao (2010-2013) Worked with laser tweezers to study long loop G-Quadruplexes and studied the kinetics of various minor species of G-Quadruplexes.

Awards/Honors/Fellowships:

Addison-Wesley Award in Physics
John Wiley Award in Physics
Choose Ohio First Scholarship

Publications (published/in press) [*indicates equal contribution]

9. **Bohrer CH**, Yang X, Weng X, Tenner B, Ross B, Mcquillen R, Zhang J, Roberts E, Xiao J. A Pairwise Distance Distribution Correction (DDC) algorithm for blinking-free super-resolution microscopy. BioRxiv. 2019 Jan 1:768051.
8. Weng X*, **Bohrer CH***, Bettridge K, Lagda AC, Cagliero C, Jin DJ, Xiao J. Spatial organization of RNA polymerase and its relationship with transcription in Escherichia coli. Proceedings of the National Academy of Sciences. 2019 Oct 1;116(40):20115-23.
7. Fang X, Liu Q, **Bohrer CH**, Hensel Z, Han W, Wang J, Xiao J. Cell fate potentials and switching kinetics uncovered in a classic bistable genetic switch. Nature communications. 2018 Jul 17;9(1):2787.
6. **Bohrer CH***, Yang X*, Lyu Z, Wang SC, Xiao J. Improved single-molecule localization precision in astigmatism-based 3D superresolution imaging using weighted likelihood estimation. BioRxiv. 2018 Jan 1:304816.
5. Klein M, Sharma R, **Bohrer CH**, Avelis CM, Roberts E. Biospark: scalable analysis of large numerical datasets from biological simulations and experiments using Hadoop and Spark. Bioinformatics. 2017 Jan 15;33(2):303-5.
4. **Bohrer CH**, Bettridge K, Xiao J. Reduction of confinement error in single-molecule tracking in live bacterial cells using SPICER. Biophysical journal. 2017 Feb 28;112(4):568-74.
3. **Bohrer CH**, Roberts E. A biophysical model of supercoiling dependent transcription predicts a structural aspect to gene regulation. BMC biophysics. 2016 Dec;9(1):2.
2. Roberts E, Be'er S, **Bohrer CH**, Sharma R, Assaf M. Dynamics of simple gene-network motifs subject to extrinsic fluctuations. Physical Review E. 2015 Dec 31;92(6):062717.
1. Koirala D, Ghimire C, **Bohrer CH**, Sannohe Y, Sugiyama H, Mao H. Long-loop G-quadruplexes are misfolded population minorities with fast transition kinetics in human telomeric sequences. Journal of the American Chemical Society. 2013 Jan 31;135(6):2235-41.

Other Publications (submitted/in preparation)

1. Christopher H. Bohrer, Jie Xiao. (2020, Review Article) "Complex Diffusion within Bacteria," In preparation.
2. Ryan McQuillen, Christopher H. Bohrer, Xinxing Yang, Jason Lyu and Jie Xiao. (2020) "Rapid light-triggered spatial reorganization of proteins inside living bacteria cells." In preparation.
3. Xiaoli Weng, Christopher H. Bohrer, Arvin Lagda, Sankar Adhya and Jie Xiao. (2020) "The spatial relationship of rrn operons in E.coli and its functional significance" In preparation.
4. Brian Tenner, Brian Ross, Dan Ohadi, Christopher H. Bohrer, Jie Xiao, P. Rangamani and Jin Zhang (2020) "Spatially compartmentalized phase regulation in the Ca2+-cAMP-PKA oscillatory circuit." In preparation.